Fall 2014

Embodied interaction with visualization and spatial navigation in time-sensitive scenarios

Yu-Ting Li
Purdue University

Graduate School Form 30
(Updated 11/20/2014)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Yu-Ting Li

Entitled

Embodied Interaction with Visualization and Spatial Navigation in Time-Sensitive Scenarios

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Juan Wachs

Eugenio Culurciello

Shimon Nof

Brad Duerstock

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Juan Wachs

Approved by Major Professor(s): _____

Approved by: Abhi Deshmukh 12/05/2014

Head of the Department Graduate Program Date

EMBODIED INTERACTION WITH VISUALIZATION AND SPATIAL NAVIGATION IN TIME-SENSITIVE SCENARIOS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Yu-Ting Li

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2014

Purdue University

West Lafayette, Indiana

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Juan Wachs for the advisory and support with his patience and knowledge during my Ph.D. study and research.

My sincere thanks also goes to the rest of my thesis committee: Dr. Shimon Nof, Dr. Bradley Duerstock, and Dr. Eugenio Culurciello for their encouragement and helpful comments.

I would like to acknowledge the financial, academic support of US Air Force Office of Scientific Research (AFOSR), and Dr. Paul Havig, who provides guidance and expert opinion to my research.

I also thank my labmates in ISAT Lab: Mithun Jacob, Hairong Jiang, Tian Zhou, Maru Cabrera, and Ting Zhang, and all my dear friends: Betty Kuo, Lilian Lin, Tiffany Lin, Winni Chen, Annen Chen, RainSwl Chang, and Yun-Shu Wang.

Last but not the least, a special thank to my family. I would like to thank my parents, sister, and grandparents for always being there with me while I was so far away from home. Without your support and love, I could not have completed this Ph.D.

TABLE OF CONTENTS

Page

LIST OF TABLES v
LIST OF FIGURES vi
LIST OF ABBREVIATIONS viii
LIST OF SYMBOLS x
ABSTRACT…… xii
CHAPTER 1. INTRODUCTION 1
1.1 Background 1
1.2 Research Problem 4
1.3 Research Questions 5
1.4 Research Contributions 6
1.5 Summary 6
CHAPTER 2. LITERATURE REVIEW 8
2.1 Embodied Interaction 8
2.2 Attention 10
2.2.1 Selective Attention 12
2.2.2 Attention Control 14
2.2.3 Attentive User Interfaces 15
2.3 Human Computer Interaction 19
2.4 Interaction Modalities 22
2.4.1 Gesture Recognition 23
2.4.2 Speech Recognition 25
2.4.3 Foot gesture 26
2.5 Feedback 27
2.6 Knowledge of Results 28
2.7 Bayesian Network 29
CHAPTER 3. METHODOLOGY 34

Page

3.1 System Architecture 36
3.2 Spatial Navigational Task 38
3.2.1 Traveling salesman problem 38
3.2.2 Cyber-Physical Navigational Problems 40
3.3 Using Embodied Interaction for a Spatial Navigational Task 41
3.4 Bayesian Attentional Network 44
3.4.1 Determining the BAN Structure by Operator's Knowledge 49
3.4.2 Determining the BAN Structure through Evolutionary Learning 50
3.5 Consensus (Majority) Model 54
3.6 Utility-Directed Feedback Model 57
3.6.1 Cost and Benefit Metrics 57
3.6.2 Expected Utility Function 59
CHAPTER 4. EXPERIMENTAL RESULTS 61
4.1 Case Study 1 61
4.1.1 Design of Experiments 62
4.1.2 Results: Bayesian Attentional Networks 66
4.1.3 Enhanced Interaction Modality 71
4.1.4 Task Performance of Interaction and Feedback Modality 73
4.2 Case Study 2 77
4.2.1 Design of Experiments 81
4.2.2 Results: BAN measure vs. Secondary Task measure 82
4.2.3 Results: Modalities Usage of the Multimodal Interface 83
4.2.4 Results: Multimodal Interface vs. Keyboard Interface 85
CHAPTER 5. DISCUSSION 88
5.1 Discussion: Bayesian Attentional Network 89
5.2 Discussion: Task Performance 93
CHAPTER 6. CONCLUSIONS AND FUTURE WORK 96
REFERENCES 100
APPENDIX QUESTIONNAIRE 112
VITA 115
PUBLICATIONS 116

LIST OF TABLES

Table Page

Table 3-1 Definition of discrete states of each variable 48

Table 3-2 Example of the values of Xi 53

Table 4-1 Summary of collected trials for each scenario. 62

Table 4-2 Summary of collected trials for each interface 81

Table 4-3 Statistical summary of equivalence tests for attention measure 83

Table 5-1 Summary of attention-supported user interface and comparisons with our work 92

Table 5-2 Summary of previous embodied interaction based interface and comparisons with our work 94

Table 6-1 Summary of two Case Studies and their relation to the research questions 98

LIST OF FIGURES

Figure Page

Figure 2-1 Bayesian network graph for the causes of nausea; food poisoning and flu may cause nausea. 30

Figure 2-2 An example of a directed acyclic graph. 33

Figure 3-1 Framework for the study of embodied interaction through navigational experimentation, computational modeling, and decision-making. 37

Figure 3-2 A 8-city TSP with reward at each city. 40

Figure 3-3 A cyber-physical network: network latency of US Air Force Bases. 41

Figure 3-4 Five modalities used in the experiment. (a) gross gestures (Kinect) (b) fine gestures (glove) (c) speech (d) feet on dance pad (e) body stance on Wii balance board 43

Figure 3-5 Kinect skeleton of a user. 44

Figure 3-6 System architecture representing construction of the BAN. 46

Figure 3-7 Node structure as bit representation $x12x13x23$ 50

Figure 3-8 Example of crossover and mutation operations 51

Figure 3-9 The adjacency matrix of the representative BAN for the 10 candidate BANs. 56

Figure 4-1 The reward of 8 cites discounted over time 63

Figure 4-2 Experimental apparatus 64

Figure 4-3 Visualized TSP displayed to the users 65

Figure Page

Figure 4-4 Bayesian Attentional Network's structure obtained by (a) – (e) Operator based, (f) – (j) Evolutionary learning. 67

Figure 4-5 The adjacency matrix of BANs obtained by (a) – (e) Operator based, (f) – (j) Evolutionary learning 69

Figure 4-6 Representative BAN and its enhanced adjacency matrix 69

Figure 4-7 Convergence characteristics of 5 evolutionary BANs. 70

Figure 4-8 Boxplot of 10 interaction scenarios 73

Figure 4-9 Group means comparison 73

Figure 4-10 Expected utility vs. benefits and costs 76

Figure 4-11 Cyber operation tasks: layer 1 78

Figure 4-12 Cyber operation tasks: layer 2 79

Figure 4-13 Cyber operation tasks: layer 3 80

Figure 4-14 A prototyped multimodal interface used in a cyber-physical system 80

Figure 4-15 Example of stimulus sequence of a 1-back task and its correct responses for each T-like visual stimulus representation. 82

Figure 4-16 Level of assessed attention using two approaches 83

Figure 4-17 3D plot of inferred attention vs. the percentage of each modality used. 84

Figure 4-18 2D plot of inferred attention vs. the percentage of speech used 85

Figure 4-19 Expected utility vs. four performance metrics 86

Figure 4-20 Comparison of error rate between two interfaces 87

LIST OF ABBREVIATIONS

Abbreviation		Page
AAS	Attention aware system	15
AUI	Attentive user interface	15
BAN	Bayesian attentional network	35
BN	Bayesian network	29
CPD	Conditional probability distribution	29
DAG	Directed acyclic graph	29
DS	Feet movement on dance pad as interaction modality and sound as feedback modality	62
DV	Feet movement on dance pad as interaction modality and visual as feedback modality	62
EC	Embodied cognition	1
EM	Expectation maximization	52
GP	Genetic programming	50
GPS	Global positioning system	20
GS	Fine hand gesture as interaction modality and sound as feedback modality	62

Abbreviation Page

GV Fine hand gesture as interaction modality and visual as feedback modality……...……..62

GUI Graphical User Interfaces……..77

HCI Human computer interactions……11

HMM Hidden Markov model……...24

KR Knowledge of results…….28

KS Gross hand gesture as interaction modality and sound as feedback modality…….62

KV Gross hand gesture as interaction modality and visual as feedback modality…….62

MDL Minimum Description Length……33

CMM Consensus (Majority) Model…….54

RANSAC Random Sample Consensus……...55

SDK Software development kit……..41

SS Speech as interaction modality and sound as feedback modality……..62

SV Speech as interaction modality and visual as feedback modality……..62

TSP Traveling salesman problem……..38

WS Body Stance on Wii Balance Board as interaction modality and sound as feedback modality……..62

WV Body Stance on Wii Balance Board as interaction modality and visual as feedback modality……..62

LIST OF SYMBOLS

Symbol		Page
$\mathcal{A}$	The adjacency matrix of the representative BAN	55
a_{ijk}	The parameter of a given Bayesian network with Dirichlet distribution	53
B	A Bayesian network	29
B_i	Design benefit associated to performance metric i	58
C_i	Design cost associated to performance metric i	58
D_l	Dataset used for Case Study 1 collected in the experiment	48
$\boldsymbol{e}$	Observed evidences	60
e_i	Updated states of other variables X_i	32
F_j	Rendering feedback with modality j	59
$\mathbb{G}$	A directed acyclic graph	29
H	Latent variable	52
I_k	Interaction modality k	59
λ_i	The states of variable X_i	47
M	Number of observations assigned to D_i	48
M_{ij}	The number of instance in the data where X_i's predecessors are equal to the j-th possible value; that is, $M_{ij} = \sum_{k=1}^{r_i} s_{ijk}$	53
$Pa(X_i)$	Parents of X_i	31

Symbol Page

π_v Reward value allocated at vertex v……....39

Ψ A feature vector of λ_i……48

r_i The number of possible values of discrete variable X_i……..52

q_i The total possible number of different values for the set of predecessors,X_i ……52

S Raw instances collected by the sensors……47

s_{ijk} The number of samples in which X_i is equal to k and X_i's predecessors are equal to the j-th possible value……...53

Θ Conditional probability distribution of the Bayesian network……...29

$\{\theta_1, \theta_2\}$ The binary values of variable X_i……....47

T_m The maximum time allotted for visiting the current city……...39

t_v The spending time of visiting vertex v……..39

U Overall utility obtained by measuring the performance metrics……59

V Sets of observable variables……...52

v Index of the vertex in TSP……39

ω_i Weighting factor assigned to performance metric i……...60

X_i Node i of the Bayesian network……31

x_{ij} Connection between node X_i and node X_j ……50

ABSTRACT

Li, Yu-Ting. Ph.D., Purdue University, December 2014. Embodied Interaction with Visualization and Spatial Navigation in Time-Sensitive Scenarios. Major Professor: Juan P. Wachs.

Paraphrasing the theory of embodied cognition, all aspects of our cognition are determined primarily by the contextual information and the means of physical interaction with data and information. In hybrid human-machine systems involving complex decision making, continuously maintaining a high level of attention while employing a deep understanding concerning the task performed as well as its context are essential. Utilizing embodied interaction to interact with machines has the potential to promote thinking and learning according to the theory of embodied cognition proposed by Lakoff. Additionally, the hybrid human-machine system utilizing natural and intuitive communication channels (e.g., gestures, speech, and body stances) should afford an array of cognitive benefits outstripping the more static forms of interaction (e.g., computer keyboard). This research proposes such a computational framework based on a Bayesian approach; this framework infers operator's focus of attention based on the physical expressions of the operators. Specifically, this work aims to assess the effect of embodied interaction on attention during the solution of complex, time-sensitive, spatial navigational problems. Toward the goal of assessing the level of operator's attention, we present a method linking the operator's interaction utility, inference, and reasoning. The

level of attention was inferred through networks coined *Bayesian Attentional Networks* (BANs). BANs are structures describing cause-effect relationships between operator's attention, physical actions and decision-making. The proposed framework also generated a representative BAN, called the *Consensus (Majority) Model* (CMM); the CMM consists of an iteratively derived and agreed graph among candidate BANs obtained by experts and by the automatic learning process. Finally, the best combinations of interaction modalities and feedback were determined by the use of particular utility functions. This methodology was applied to a spatial navigational scenario; wherein, the operators interacted with dynamic images through a series of decision making processes. Real-world experiments were conducted to assess the framework's ability to infer the operator's levels of attention. Users were instructed to complete a series of spatial-navigational tasks using an assigned pairing of an interaction modality out of five categories (vision-based gesture, glove-based gesture, speech, feet, or body balance) and a feedback modality out of two (visual-based or auditory-based). Experimental results have confirmed that physical expressions are a determining factor in the quality of the solutions in a spatial navigational problem. Moreover, it was found that the combination of foot gestures with visual feedback resulted in the best task performance ($p < .001$). Results have also shown that embodied interaction-based multimodal interface decreased execution errors that occurred in the cyber-physical scenarios ($p < .001$). Therefore we conclude that appropriate use of interaction and feedback modalities allows the operators maintain their focus of attention, reduce errors, and enhance task performance in solving the decision making problems.

CHAPTER 1. INTRODUCTION

1.1 Background

We live in a period of time when both the mobility and ubiquity of computing devices make it possible for end-users to access and interact with information at any moment and virtually anywhere on the globe. Amongst these numerous computing devices are: PCs, laptops, smart phones, and tablets. These devices are commonly encoded with methods of interaction by which humans can relate and manipulate the devices. The ability of these devices to convey information to users and to perceive new/unstructured environments is not only determined by the overall perceived experience of the user, but also the degree of efficacy by which tasks are accomplished. In this context, a key factor in achieving effective utility on these devices consists of adapting suitable modalities of interaction according to the level of attention required to complete a successful task. Finding the physical actions (required to operate the devices) that can offer relative advantage in terms of problem solving when compared to traditional methods of interaction is also vital. It has been indicated that traditional interfaces are limited when used to complete tasks associated with complex data visualization and navigation in information spaces [1], [2]. Current trends in complex image analysis and visualization involve using more of the human body [3], [4], rather than a more passive form of analysis (e.g., users seated in front of computer screens). This trend is well rooted within Embodied Cognition (EC)

theory, which maintains that all human cognition is shaped by aspects of the human body [5]. An example where embodied interactions have shown clear advantages over traditional forms of interaction is during interactions with overhead imagery. Such forms of interaction would allow for analysts and operators to maintain their attention on the imagery and use more of their body while performing analytic tasks; this also helps to eliminate the need for functional navigation menus traversed via keyboard or mouse. Such analytic tasks are the brick and mortar for several decision making processes, such as those existing in medical experts systems, air traffic control systems, and cyber-physical systems. Furthermore, decision making and action are deeply integrated as people are usually processing their actions during task completion. Part of human reasoning is influenced by physical interaction with the environment just as bodily activities are affected by thought [6]. The claims surrounding this relationship have been tested in a number of experiments related to Visual Search [7]; Distance Perception [8]; Language Processing [9]; Memory [10]; Science Education, [11], [12] and Performing Arts [13], [14].

This dissertation investigates the use of embodied interaction to support spatial optimization for navigational problems through the rigorous use of mathematically, biologically and psychologically-inspired methods. These methods include: a systematic characterization of the operator's physical interactions with the machine while solving complex spatial navigational problems; probabilistic modeling of the links between attention and task performance; evolutionary inspired approaches for network generation; and the development of metrics based on utility theory to assess and derive suitable interaction and feedback modalities. To validate such a framework, we conducted two

real-world experiments with visual interaction systems designed in two stages. Case Study 1 involved a systematic characterization of the operator's physical interactions while solving a spatial navigational problem (i.e., Traveling Salesman Problem) while they are permitted to navigate through visual representations of the problems. The best combination of interaction and feedback modalities was determined for this time-sensitive, and dynamic decision making scenario. Case Study 2 was designed for the visualization of cyber-operations during which operators interact and navigate through datasets of cyber-physical visual information to resolve cyber threat using multimodal interactions.

Several terms employed throughout this dissertation are herein defined:

(i) Interaction modality – a communication channel which enables an operator to interact with the system.

(ii) Feedback –a message passed from the system to the operator with the express purpose of informing the user about the current state of the system.

(iii) Command – a directive to the system to perform a specific task.

(iv) Lexicon – the list of assigned commands to particular interaction modalities.

(v) Primary task – the main task assigned to a user, this task takes priority over all others.

(vi) Secondary task – a peripheral task that is conducted simultaneously with the primary task.

(vii) Dual-task – the scenario of completing two distinct tasks at the same time.

1.2 Research Problem

Embodiment offers various cognitive advantages, including better information retention, context comprehension, and mathematical reasoning. Those advantages are particularly important during complex problem solving. The research problem that we are trying to address is determining whether or not embodied interaction leads to better decision making in spatial navigational problems, which has not been quantitatively proven yet. The objective of this dissertation is to propose an analytic framework to determine the suitable physical expressions performable by the human body that lead to enhanced spatial navigational problems solving. This is done through the combination of Bayesian theory, evolutionary based methods, and utility functions.

Traditional interfaces are limited in dealings with complex applications or spatial data and thusly, are not the most suitable interfaces for the navigation of various visual and data analysis environments. The proposed framework provides a solution for determining whether embodied interactions lead to better decision making; it involves analytically determining the most suitable combination of control and feedback modalities. This combination may eventually lead to any of the following: a reduction in the cognitive burden on the user; an enhancement of his/her performance; and better decision-making while performing in complex settings. A natural, human-centered, multimodal interface may additionally enable operators by allowing them to utilize multiple types of communications (e.g., hand gestures, speech, foot movements); this embodied interaction based multimodal interface is herein tested, and compared with a non-embodied interaction based interface using the proposed framework in hopes of observing such benefits. The goal of this research is to propose a method by which one may support

spatial navigational problem solving through the use of a natural, and multimodal interaction.

1.3 Research Questions

There are two research questions (RQs) studied in this dissertation:

RQ1: What is the optimal combination of interaction modalities and feedback that lead to the best task performance (among the alternatives studied)? Within the context of this thesis better task performance means, higher accuracy, shorter completion time, and improved quality in performing a navigational task.

To specifically address RQ1, this research work shall explore different combination of interaction and feedback modalities used during a navigational task and determine the optimal combination by determining which has significantly better task performance metrics. The answer to RQ1 is expected to be the combination of interaction and feedback modalities with higher accuracy, shorter completion time, and improved quality statistically.

RQ2: Which benefits are offered by embodied interaction over those offered by non-embodied interaction method during the completion of spatial navigational scenarios?

Embodiment offers cognitive advantages such as information retention and problem reasoning. In contrast, non-embodied interaction based interfaces create a gap between a user's intent and the execution of the intent. In RQ2, the potential advantages offered by embodied interaction over those offered by non-embodied interaction method are expected to include improved quality of solution, and lower execution errors.

1.4 Research Contributions

Research in the area of embodied cognition has shown that physical interaction, attention and task performance are closely related [6]. No studies have been found to provide analytical methods for expressing and developing this relationship. Understanding this relationship would allow for the creation of new and more effective means of performing complex image analysis and visualization through the use of our bodies. This is supported by the EC theory, which states that all human cognition is shaped by aspects of the human body [5]. This dissertation shall focus specifically on the embodiment principle as a means of interaction with visual and spatial information. This is done in order to explore the effect of embodiment on complex decision making. Herein, we present two main contributions: an evaluation of the operator's level of attention by building a cause-effect relationship between physical actions, task performance, and level of attention; and a determination of the most suitable combination of interaction and feedback modalities so as to enhance the operator's decision-making (e.g., higher accuracy, shorter completion time, and improved quality). To the best of our knowledge, this is the first time that this relationship has been established and studied, as well as the first time the effects of this relationship were reported through systematic analysis.

1.5 Summary

This chapter has provided an overview of the intents and methods of our research. We have included: a background on the problem; definition of key terms; research problem; research questions; and our research contribution. The remainder of this document is organized as follows: Chapter 2 summarizes an overview of previous research related to

several key areas of this work; Chapter 3 introduces the system architecture and proposed framework; Chapter 4 presents the case studies conducted by real-world experiments validating the methodology proposed in Chapter 3. Chapter 5 discusses the results and findings; finally, conclusions and future work is presented in Chapter 6.

CHAPTER 2. LITERATURE REVIEW

This chapter should act as an overview of the current research related to our study. This chapter begins with the description of basic concepts and previous research related to embodied interaction, focus of attention, multimodal interfaces, and feedback, as well as a description of the main technological advances in each component employed within the system presented.

2.1 Embodied Interaction

Embodied interaction theory has been studied and comprises the user's senses, the environment, and information acquisition; all of these are combined by an operator to show intention by means of physical action [15]. This concept is closely linked to the concept of embodied cognition from psychology [5]. Embodied cognition, as a theory, postulates that our cognition is affected by our interactions with our environment. This implies that the environment plays an important role both during cognitive processes and in the formulation of our cognitive processes [16], [17].

Embodied interaction has been shown to promote both thinking and learning [18]–[20]. Segal [19] showed that the use of tactile-enabled digital devices yields better performances from their operator in simple algebraic operations when compared to traditional interfaces (such as monitor, keyboard, and mouse). In Segal's experiment,

young children were instructed to perform tasks comprised of counting, addition, and estimating numbers on a numerical axis. The children who were allotted gestural interfaces that integrated higher levels of behavior mapping and direct touch outperformed children who were apportioned traditional interfaces. It has also been shown that gesture is capable of triggering mental images which may help in the solving of spatial-visualization problems [21]. Chu and Kita [21] investigated the beneficial role of gestures by conducting experiments concerning the mental rotation and paper folding tasks. The use of gestures were spontaneously aroused from the participants who previous exhibited difficulties in solving spatial-visualization problems, consequently performance was improved. All these studies aided in the confirmation that embodied cognition is not a unitary theory; it utilizes various human capabilities involving: motor control, focus of attention, visual perception and spatial cognition [22]. Discussed in this dissertation, we designed and employed a study requiring the users to solve a spatial-navigational problem while allowing them a variety of body actions as a form of communication. This would eventually stress specific cognitive benefits associated with embodied interaction in problem solving and decision-making.

Embodied interaction also relates to the ways by which people interact mentally and physically with information technology; in this manner, it has been considered to be a novel approach in human computer interactions (HCI) [15] for both information visualization and navigation. As people often employ their bodies to picture and to describe both images and ideas in mind, it is understandable that gestures are widely used to express spatial and motor information [23]. This may explain why complex image analysis and visualization can often benefit from the engagement of the human body

during both interaction with and analysis of visual information [24]–[27]. However, current trends within information visualization systems still rely on the use of the keyboard-and-mouse pair for interaction with data. These traditional interfaces are limited when users deal with either complex applications or spatial data (especially multi-dimensional data)[1], [2]. Furthermore, these interfaces have been shown to be less suitable for a user interfacing with visual-and-analysis type of environments (e.g., meeting room, public spaces, and operating room)[28], [29]. Another limitation of traditional interfaces is the organization of objects within hierarchical navigation structures, where options are only found within folders and drop-down menus; thus users may become frustrated by having to assess and search through layers of options [30]. Van Dam [30] has also shown that traditional interfaces also introduce significant "cognitive distance", this is to say that a gap is generated between the operator's intent and the execution of said intent.

Embodied interaction have been shown to enhance thinking and learning so as to improve task performance. The performance of an assigned task is also greatly affected by whether or not the user is correctly focusing their attention on the task [31]. The following chapter will review the literature about focus of attention.

2.2 Attention

Cognitive psychologists refer to attention as the cognitive process of selectively concentration on processing only specific information related to the subject's intended environment or task while ignoring other information. According to psychologist William James, attention is, "the taking possession by the mind, in a clear and vivid form of one

out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration of consciousness are of its essence. It implies withdrawal from some things in order to deal with others" [32]. Attention is a limited resource, thus it is considered to be the most valuable and scarcest commodity in the context of human computer interaction (HCI) [33]. A capacity model of attention assumes that due to the limited capacity available for performing mental work, some mental activities have different demands on this limited resource than do others [34]. The multivariate nature of attention quantification presents an especially difficult challenge for quantitative, objective and evidence-based measurements [35]. There exists other more indirect methods by which to assess attention; these rely upon the basic paradigm of dual-task performance. An example of these is to observe the ability/performance of a subject in resolving a number of system alerts while performing a specific main task. It was found that the more alerts that the user can resolve without affecting task performance is a good indication of the level of attention which the user may allocate to the primary task [36]. A degradation in task performance under these circumstances is referred to as “interference”. The most accepted explanation of dual-task interference relies on the fact that cognitive resources are both finite and shareable; since each presented task is tapping the same cognitive resources, an increase in consumption due to the primary task lowers the resource available for any secondary task. [37]. The degree of interference is heavily affected by facets such as: the similarity of tasks; difficulty of tasks; or the level of practice or expertise [38].

However, performance-based measures possess their own drawback. The main drawback is that it is difficult to isolate the effects of attention from the decision-making process

based only on the operator's performance [39]. Errors associated with this interference may actually reflect inaccurate decision making; though inaccurate decision making may reflect effective focus of attention, there exist a variety of other possible causes, such as errors in judgment, misinformation or a lack of contextual knowledge.

Nevertheless, assuming users operating under equal contextual conditions (same ability to make judgments, same knowledge, etc.), focus of attention has a decisive effect in the decision-making process, and that is why determining it is key to infer task performance. Other approaches attempting to determine focus of attention rely on physiological signatures (e.g., ocular movement [40], [41] or heart rate [42]) require the operator to stay seated, such that the acquired physiological signals are not masked by physical forms of interaction. Assessing focus of attention using physiological signatures while the body is in motion is still an open research question. This dissertation will not rely on physiological signatures to assess focus of attention due to the issues described earlier, and instead will follow the performance-based measures as the baseline to assess attention.

2.2.1 Selective Attention

Selective attention refers to the differential processing of information from disparate yet simultaneous sources [43]. Note that the varied sources of information can be either internal (memory and knowledge) or external (objects and events in the surrounding environment). The process of filtering out the information from these less desirable stimuli while still extracting supplemental information from the stimuli on which the attention is focused is vital for completion of differential processing. Selective attention

theories suggest that people place a consistent emphasis on a class of perceived objects or events from which information is processed in preference to other sources. The cocktail party effect, described in [44], is an example of this theory; in this example people can selectively listen to a conversation in which they may be interested while remaining inattentive to other surrounding conversations. This effect illustrates that a person is capable of focusing his/her auditory attention on a particular stimulus while filtering out all other stimuli. Cherry [44] conducted the experiment under various conditions in which two different messages were mixed and played to both ears (diotic) as well as unmixed and played to different ears (dichotic); the subjects were then asked to repeat one of the two messages by both speaking it out and writing it verbatim. The findings showed that the primary message was received by the subjects while the secondary message was not received. On the contrary, Johnston and Dark [43] found that divided attention occurs when a speaker who is engaged in a conversation (in the context of the cocktail party example) notices salient important information (for example, his own name) that is part of a simultaneous occurring different conversation. Divided attention is a process of integrating multiple parallel stimuli by allocating the available attention-based resources on more than a single task at a given time [45]. In order to conduct two simultaneous tasks (as previously discussed), divided attention is required. A common example of dual-task scenario, is simultaneously driving and conversing on a cell phone [46]. In this particular case, performing a secondary task – conversing on a cell phone – while driving has been shown to have a negative effect on the ability to complete the primary task, driving [47]–[49]. Potential competition for cognitive resources occurs in such dual-task scenarios; this provides an explanation why – in the driving example – there is a higher

incidence of errors on complex routes, such as collision, sudden braking, and missed turns [46].

2.2.2 Attention Control

There exist two common mechanisms by which the attention is aroused; these being bottom-up processing, and top-down processing [50]. Bottom-up processing is known as stimulus-driven attention, or exogenous attention; in this mode of processing, attention is driven by salient stimuli such as the flashing of a fire alarm. In top-down processing, the stimulus is derived from knowledge concerning the necessity of current task such as finding a lost key. Thus, top-down processing is sometimes called goal-driven or endogenous attention [51]. Endogenous attention processes are voluntary, effortful and sustained; while, exogenous attention is transient and attention is attracted to its source automatically [52]. An example of an exogenous attention process (bottom-up mechanism) is when we visually observe an environment, some objects with significant features (such as color, shape, and orientation) draw our attention automatically. Previous research has focused on explaining human visual attention as being a result of a bottom-up process mechanism [53].

Within this dissertation, we operate under the assumption that the operator has intention and knowledge while performing the presented task, and is attempting to achieve the goals associated with the task. The classification of the arousal of attention is considered the top-down mechanism within our works presented herein.

2.2.3 Attentive User Interfaces

Determining the necessary cognitive resources (e.g., user's focus of attention) required to operate and interact effectively with a given computing device is a central question that human factor engineers need to frequently address during a product life cycle [54], [55]. Working in this context, Attentive User Interfaces (AUIs) are a class of those computing interfaces which are designed to be sensitive to the user's level of attention; AUIs therefore offer the ability to support the user's attention goals through certain features in their design [56]. AUIs have also been referred to as Attention Aware Systems (AAS) [57]. Generally speaking, AUIs have been used to track user's goals and level of attention on a given computer-based task. Given this, little research has been conducted evaluating the cognitive cost expended during switching between competing scenarios with the same attentional resources [58], [59]. Within these works AUIs were designed to support the user's attention-based processes such that the user's focus of attention is allocated efficiently during the completion of a task. This concept is instrumental in the design of portable wireless computing devices [60].

With the modern ubiquitous nature of computing access, people are capable of carrying/wearing multiple computing devices, such as smartphones, laptops, tablets, and head-mounted displays (e.g., Google Glass [61]) at any time, almost everywhere. Most of these devices are interconnected by wireless networks and therefore both the timing and the means of transferring relevant information between such device and finally to the user may lead to repetitive interruptions in the user's attention [62]. In this context, the goal of AUIs is to effectively measure the priorities of the user so that their resource of attention is allocated optimally to the assigned and desired task possessing the highest priority.

Given the user's current task and relevant aspects of their presented immersive environment, the system must observe the sensory cues expressed by the users to obtain information concerning their current activity and the state of the environment. Such systems must be able to detect possible alternative foci that should be permitted to the user given the current level of attention. As an example, given the state of attention of a user, the AUI must evaluate whether or not an alert related to a new email in the user's inbox should be immediately presented [57]. With this example the AUI must consider not only the timing of the alert, but also its form. The cost/effectiveness of alternating possible focus should be also considered. After the above processes are completed, strategies for information presentation are determined, such as the modality, the content presented, and the timing of the presentation [57].

There are five key properties for AUIs [63]:

1. "Sensing attention": determining to which machine, user, or job the user is most likely paying attention to at a given instant.
2. "Reasoning about attention": modeling the user's interactive behavior in order to understand user's task prioritization.
3. "Communication of attention": conveying information about the user's attention to the agents within their environment.
4. "Gradual negotiation of turns": determining user's availability for and the appropriateness of an interruption.
5. "Augmentation of focus": emphasizing the information desirable for focus while attenuating any peripheral details.

One of the original AUIs was Rick Bolt's Gaze-Orchestrated Dynamic Windows [64] of the early 80s, which employed gaze tracking as a means to interact with computers. The system sensed the user's visual focus of attention within an image and then increased/enhanced the foci upon a large display. In late 90s, Jacob [65] and Zhai et al. [66] showed that an eye tracking system could be used to assess the user's attention. The gaze-responsive self-disclosing display [67] monitored the user's gaze behavior and responded accordingly with a comment. iTourist [68] was developed for city trip planning, which exploited the user's gaze pattern to provide further information about a city of interest. GAZE [69] used eye trackers to assess visual awareness among users in a group and leverage this information to mediate effective communication and collaboration within that group. In GAZE-2 [70], gaze-awareness was leveraged to support group video conferencing. Regions showing high visual interest within a group of users were zoomed to facilitate better interactions with the group. Current cutting edge technology that employs gaze tracking has been incorporated into consumer products. The Samsung Galaxy S4 [71] is equipped with the front-facing camera which is used to automatically pause video if the user's eyes have deviated from the screen, and conversely, resume the video once the user's gaze return to the image.

There is some limited work concerning the use of Bayesian models for AUIs to determine the most appropriate form for the delivery of alerts to the user. An example of Bayesian-based AUIs is Priorities System [33] which employed classifiers for the prediction of the urgency of incoming emails and then to decide the most appropriate time of notification. By considering the cost of interruption and the cost of delayed review for each incoming message, a decision of whether, when, and how the alerted is transmitted was made to

minimize distraction. Horvitz et al. [72] presented a networked Bayesian forecasting service, COORDINATE, which predicts the user's presence and availability. Bayesian learning and inference have been applied in Horvitz' work to build probabilistic models of attention. The Notification Platform [73] probabilistically observed a user's level of attention based on a Bayesian attention model. That model collected the perceptual evidence (gaze, utterance) provided by the user and scheduled activities (online calendar, location sensing) accordingly. Context-sensitive cost of distraction is said to compute the expected utility of each channel and modality used to transmit the alert to the user. Decisions concerning the channel and modality with the highest expected value are solved for in an optimized manner. The benefits of alerts and the costs of deferring notification were weighed to modulate the communication of notifications to the user. Regardless of this mediation, interruptions are generated each time notifications from the devices occur. A method to infer the state of interruptability of the user and to predict expected cost of interruption was presented by Horvitz and Apacible [74]. In their work, after sensing the state of the operator's attention, the expected cost of interruption was computed given a probability distribution over attention and a utility function.

In our work, rather than observing user's scheduled activities, we chose to focus on embodied interactions (gestures, utterance, and body stance) used to manipulate, navigate and interact with visual information. This distinction is important since we admit a level of uncertainty concerning the operator's next best decision, as opposed to maintaining a schedule of activities which are planned beforehand; we allow for a level of fluidity in the designed sequences. This work differs from those of Horvitz as the natural and intuitive forms of interactions are explored in addition to the maintenance of required

focus of attention. Furthermore, user-dependent interaction is considered to be a facet which varies between users. To manage the focus of attention, the characteristics of interaction between a user and an interface are essential and will be introduced in the next section.

2.3 Human Computer Interaction

Human-computer interaction (HCI) is a discipline which involves the design, evaluation, and implementation of means by which the communication between computing systems and humans may be enhanced [75]. This field is inter-disciplinary in nature, spanning across computer science, psychology and cognitive science; human factors and ergonomics, and industrial engineering and design must all also be considered [75], [76]. From the perspective of computer science, focus is placed on the design and engineering of human interfaces and applications [77]. From psychology, the focus is on the empirical analysis of user behaviors and application of cognitive process theories. From the industrial engineering and design perspective, the focus is interactive product design. Each discipline, therefore, places the emphasis upon different aspects of HCI. The ultimate goal of each discipline is to create theories, methods, and practices that may increase the usability of computing systems and lead to higher user satisfaction and productivity [78]. The desired outcome will provide a balance among the needs of the user, the machine's capabilities, and the required services involved in the generation of both quality and optimal performance of tasks [79].

With increasing computing capabilities from emerging wearable devices and continuous access to social networks [80], interactions with mobile devices offer a new dimension

within HCI (mobile HCI) [81] that was previously unavailable. A key aspect in successful HCI design is determining the proper channels (or modalities) by which users are permitted to interact with computing systems. For example, touch screen based interaction is becoming almost a standard mode of interacting with mobile devices [82]. Voice recognition is becoming more popular for accessing global positioning systems (GPS) while the user simultaneously drives a car [83], interacting with on-demand internet television [84], or mobile based services like Siri [85]. Haptics is another common method to deliver alerts in noise restricted situations [86] and it is commonly found in cellular notification of text and other messages.

Based on the nature of the communication, there are three main modalities for the communications and interactions; these modalities are defined within the following categories [79]:

1. Vision-based
2. Audio-based
3. Sensor-based

Vision-based HCI deals with human responses that are clearly expressed through some type of user activity captured through a color or depth sensor. Activities such as hand gestures, body movements, facial expressions, and gaze detection [87] are common forms of user interaction. In audio-based interaction, emphasis is placed on using speech and natural language to interact with devices. Included in the audio-based category are: speech recognition, utterance detection, auditory emotion analysis, and musical interaction. The last category, sensor-based interaction (excluding optic and acoustic sensors which belong to other categories), is very diverse with a range of applications

such as manipulating an input/output sensor mediating between the user and computer interface [79]. Examples of this class of devices includes: computer mice, keyboards, joysticks, haptic sensors, and pressure sensors [88]. Among these presented interaction channels, the most commonly used for interaction are keyboards and mice (which are a simple form of sensor-based interaction). Novel approaches – including gesture, speech, haptics, eye movement and blinks – are emerging modalities of interaction stemming from the development of new hardware and software which allow for precise and fast recognition of these input/output modes [89]. The overall trend is to facilitate user interactions with computing devices by providing more natural, intuitive, and expressive interfaces [90]. Thus, an interesting research goal is to find more accessible, intuitive, and less constrained methods to incorporate aspects of human interpersonal communication into HCI, and to develop the necessary scientific tools to attain this goal [91]. An area which may not have received enough attention is the development of quantitative and analytical tools with which to assess the effectiveness of the different interaction modalities [92], [93]. Since the best modality for an application depends on the context of use, new methodologies must consider the environment in which they are employed and capabilities of the user [94]. The works contained within this dissertation aim to bridge the gap between the user and interface by providing optimal interaction modalities through the evaluation of the quantified costs and benefits which result from the interaction.

2.4 Interaction Modalities

Humans typically use five senses to perceive the environment: vision, audition, taction, olfaction, and gustation. These sensory modalities allow the user to gather information about their environment [95]. The term "modality" in attention studies usually refers to the method by which humans perceive their environment through their sense. However, this term in human-computer studies broadly refers to the channels by which users communicate with the interfaces [89]. Throughout the development of computing technologies, certain metaphors have been adopted that enable different input modalities of devices with facets corresponding to the major human senses: camera (visual), microphones (auditory), haptic sensors (touch), and olfactory sensors (smell) [96]. The effective choice of communication channel with compatible technologies is one of the principle areas of research in HCI. Several forms of user communication have been considered for potential use in interacting with devices; these include both verbal and nonverbal communication, such as gaze, gesture, and proxemics. It is also possible to use multiple modalities simultaneously or asynchronously when interacting with a device [97]. A multimodal example is when referring to an object, an individual may speak of, point at, and look at the object, simultaneously. Similarly, as humans interact with a device, interactions such as watching the screen, typing, clicking, or speaking to the microphone may be employed concurrently. People determine the most suitable form of HCI interaction based on practical, ergonomic, cognitive and technical factors [97].

In the following subsections, our discussion will focus on specific interaction modalities which involve hand and foot gesture recognition, speech recognition, and body stance detection.

2.4.1 Gesture Recognition

Hand gestures are one of the most common and important forms of non-verbal communication among people both within the same and different cultures [98]. Gesture is considered one of the most expressive, intuitive and natural form of interaction in physical and virtual environments, when interfacing between human and computing devices [99]. Hand gestures are used in HCI since they allow for a level of natural expression with relatively low cost which can rarely be achieved through existing standard interfaces [100]–[103]. Gesture detection and recognition is the field that focus on the study of how computing systems can make sense of the gestures articulated by the users [104]. This field involves the modeling, representation, analysis and interpretation of gestures; this all is based on input signals representing various attributes related to the generation, configuration, and shape, of gestures [100]. Vision-based recognition, in contrast to glove-based recognition, is the current general approach for gesture recognition since it enables the user to remain un-tethered to devices during the interaction [105].

Vision-based interaction relies on determining image cues – such as optical color and depth information found during the image acquisition phase or found after the image were processed by computer vision techniques [106]. Several methods for gesture recognition have been studied over the last few decades, including powerful features to characterize instances of gestures (e.g., integral images [107], histograms of gradients [108], geometric moments [109], contour silhouettes [110], 3D hand skeletons [111]), and robust classification methods from the pattern recognition field (e.g., Artificial

Neural Networks (ANN) [112], Support Vector Machines (SVM) [113], Hidden Markov Models (HMM) [111][112], and Markov Random Fields (MRF) [116]).

Depth acquisition is a key technology in obtaining the depth of field information contained in the trajectories of hands. To this end, methods such as stereo vision (e.g., Leap Motion), structured light (e.g., Kinect), and laser range finders are currently being used to measure the distance to the nearest physical object from the sensor. Since 2011, the most popular and affordable depth sensor is the Kinect, a device particular to the Xbox 360 console but also functional when connected directly to a computer. The Kinect yields an RGB camera for capturing the colored images (RGB) and an infrared (IR) emitter/camera pair to measure the depth (D) information. The depth measurement is accomplished by projecting a fixed pattern of infrared light and computes the distance to any point within the field of view, based on the distortions of the projected pattern. The captured RGBD data contains both the visual and the geometric attributes of an image, enabling the Kinect device to be flexible and utilizable in many areas [117]. Applications of gesture recognition using the Kinect device include: interactive display [118]–[120], robot motion control [121]–[124], and sign language recognition [125], [126].

One of the main challenges with the study of gesture interaction is recognizing gestures amidst challenging environments, uncontrolled illumination conditions, occlusion, dynamic objects in the background, cluttered background (and articulation) or distorted objects. Temporal segmentation of the gesture (determination of the start and end boundary of a legitimate gesture) is another challenge, as the occurrences of gestures vary dynamically in duration [127]. Additionally, an important requirement for the gesture recognition system is adaptability. Such systems are required to be independent of the

type of user, the user's familiarity with the system, and the user's compatibility with the system. The system should also be independent of any cognitive mapping of gestures to the commands [128].

2.4.2 Speech Recognition

Speech is the primary form of communication between humans. The development of speech recognition technology has allowed machines to detect and differentiate human language and has been applied to various applications such as in-car systems, health care, as well as mobile devices. The widely used Hidden Markov Models (HMM) has been the main foundation for automatic speech recognition since the mid 1980's [129]. Most of modern speech recognition systems are established based on HMM, which probabilistically model the variability of the acoustic signal. The artificial neural network (ANN) [130], was also introduced in the late 1980s as another approach for automatic speech recognition.

The success of speech recognition technologies within the past decades has led to several notable speech recognition software including the Sphinx system [131] developed by Carnegie Mellon University, based on HMMs. Another resource for automatic speech recognition is the Hidden Markov Model Toolkit (HTK) [132] published by a group at Cambridge University. A discussion of state-of-art techniques for automatic speech recognition are beyond the scope of this dissertation.

2.4.3 Foot gesture

Recently, foot-based HCI has gained attention in some application domains ranging from fall detection [133], [134], training simulations [135], immersive virtual environments [136], to physiotherapy [137]. The commercial successes of foot-based gaming technologies such as the Nintendo Wii Balance Board, and the dance pad (typically used in the Dance Dance Revolution games) led to the development of foot-controlled input devices for related HCI research. Additionally, foot gestures are a suitable form of interaction to convey navigation intent than that offered through hand gestures, since feet movements resemble "walking" during exploration [138]. For example, the use of hand gestures to control panning through a map requires the user to reposition the hand again when it reaches the physical boundary of the map; prolong use of these actions, and those similar, may lead to muscular-skeletal problems such as arm fatigue. Alternatively, foot gestures have the potential advantage to continuous interaction by the ability to shift the body weight to the foot corresponding to a particular input. Several studies have investigated the use of feet movements for completing navigational tasks. Pakkanen and Raisamo [139] presented different methods for operating a graphical user interface by the foot in different non-accurate spatial tasks. Schoning et al. [140] applied multi-touch hand and foot gestures to interact with spatial information on a large interactive screen. In their study, the combinations of multi-touch hand and foot gestures input provided through the Wii Balance Board lead to faster task completion because users could perform panning and zooming simultaneously. Additionally, gaze input is used to support foot interaction in the Geographic Information Systems (GIS) [138], [141], [142].

In our work, devices including Wii Balance Board and dance pad have been be integrated into the spatial navigational task.

2.5 Feedback

In HCI, the term feedback generally refers to a form of a communication from the system to the user with the purpose of confirming the current states and intentions of the user during interaction [143]. As in interpersonal communication, dialogue involves receiving information followed by providing some type of closure, negation or affirmation concerning a spoken statement, including verbal responses (e.g., acknowledgement tokens such as “uh”, “hmm”, or “yeah” [144]) and nonverbal responses (e.g. head nodding, eye blinking, or smiles) [145], [146]. A similar expectation also exists when users interact with computing devices. For example, if the user evokes a command yet does not receive any response from the system, the user may repetitively evoke the same command or find other ways try to confirm whether or not the command was recognized successfully; i.e., repeated uses of Ctrl+Alt+Del when the computer freezes. To this end, providing feedback allows the user to be effectively immersed in the interactive environment. Higher immersion allows for better rapport with the system, which has been shown to contribute to increasing user-machine performance and decreasing failure mitigation [147]. The modalities for feedback delivery can take different forms according to the interaction “channels” used: visual, audio, tactile, haptics/force, and smell [97]. The specific channel depends on the specific human-machine system with the goal of providing timely, effective and appropriate feedback considering the state of the system and user’s state [102]. For example, it has been shown that the assistance of tactile

feedback while using of mobile devices can reduce distraction and reliance on the visual channel [148]. In opposition to the clear benefits of feedback, there are cognitive and physical costs involved in the channel adopted for this interaction. From the cognitive side, the feedback presented to the user can cause distraction, and/or mask cues of either visual or acoustic [149]. From the physical side, the wearing of sensors, gloves or other devices can cause physical stress and discomfort [150], [151]. All of these may affect the user's focus of attention. Within this dissertation, we have explored the optimal feedback channel which provides capabilities emphasizing both the cognitive and the physical benefits, while reducing the physical and the cognitive costs as well as minimizing shifts in the focus of attention.

2.6 Knowledge of Results

Knowledge of results (KR) refers to the information provided to the user after her response to a stimulus, for the purpose of informing her whether she succeeds achieving a given environmental goal [152]. It is also defined as extrinsic feedback or augmented feedback, as opposed to intrinsic feedback, given to the user. An example of KR is the scores and time displayed on the scoreboard or the countdown timer which provide information about the game to the players. This information is provided as a basis for improvement on performance in the next trial. KR is viewed as one of the most important factors in the process of learning [153]. In this context, the concept of KR is similar to the term "feedback" presented in the last subsection. However, in this dissertation, we do not study the learning effect, instead we only focus on the feedback provided within a single trial. Thus in the following chapters, the term "feedback" is used to represent information

provided to the user for the purpose of informing her whether she is succeeding in achieving a given goal.

2.7 Bayesian Network

In the previous section, literature was reviewed concerning the topics of Bayesian learning and inference as part of probabilistic models used in reasoning about the attention of a user during the interaction process between the user and devices. Due to the relevance of this technique to the proposed approaches within this dissertation and based on the method's previous use in assessing attentional levels, state-of-the-art Bayesian networks will be described briefly. A Bayesian network (BN) [154], also known as a belief network, describes the probabilistic relationship between random variables and their conditional dependencies. It is a graphical model represented by the directed acyclic graph (DAG), $\mathbb{G}$, and conditional probability distribution (CPD), Θ. Thus, a BN is usually denoted as $B = (\mathbb{G}, \Theta)$. A DAG consists of a set of nodes and edges representing the random variables and their direct dependencies with a CPD describing the conditional probability distribution associated with each node and its parents. Once the graph is constructed, it allows for probabilistic inference and learning. It can be used to predict the desired variable with several possible states according to the conditional dependencies among variables. BNs have been efficiently and widely applied in real-world tasks to model causal relationships between phenomena [155]. A general example is given in Figure 2-1, it shows a causal model for the relationships between food poisoning and nausea as well as between the flu and nausea. Both food poisoning and the flu can cause nausea, but whether or not the person has food poisoning will not give any information

about the person having or not having the flu. Therefore it is assumed that the causes are independent yet may lead to the same effect.

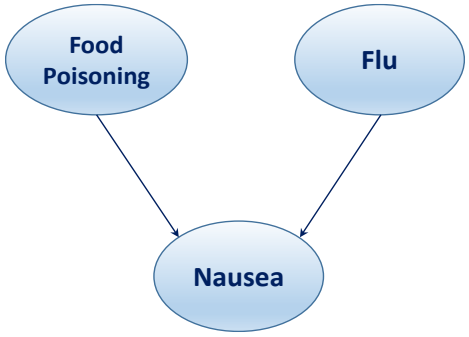


Figure 2-1 Bayesian network graph for the causes of nausea; food poisoning and flu may cause nausea.

Once the Bayesian network is constructed, the probability of certain random variables can be calculated. This probability is not stored in the model itself, but in additional tables and further calculations are needed based on the values and the topology of the network. This process is referred to as probabilistic inference in a Bayesian network [156].

Bayesian networks can be used to model a proxy for human visual attention. Example: perception is interpreted as the estimation of the posterior probability of visual features for certain target objects in an image and their locations in that image [157]–[159]. That is, developing a model for knowing what is where. The visual system aims to infer the identity and location of objects within an environment. Chikkerur et al. [159] applied the Bayesian framework to account for the search and recognition of objects in a probabilistic fashion.

Conati and Zhao [160] used Dynamic Bayesian Networks to assess students' knowledge in an educational game by tracking their actions as evidence. Sahami [161] presented a Bayesian approach for the design of an effective filter for the elimination of junk emails. To filter junk email, the Bayesian classifier was calculated from textual data included within the emails themselves. Bayesian user models were exploited to infer the software user's needs and to provide intelligent assistance to the user relying on observation of user's background, actions, and queries [162]. Gievska [163] adopted Bayesian Belief Networks (BBNs) to determine the most appropriate timing of the interruptions by the computer. The interruption mediator was found to improve task performance, support situation awareness, and allay disruption of the user's emotional state.

A Bayesian network is composed of a set of variables from a network structure, S, with the directed edges between them and a set of conditional dependencies, P, associated with each variable. Specifically, a Bayesian network includes the following:

1. A group of variables $X = \{X_1, X_2, \dots, X_N\}$ and a group of directed edges between those variables.
2. Each variable includes a finite set of mutually exclusive states.
3. A directed acyclic graph (DAG) is made of the variables and the directed edges.
4. For each variable X_i with parents $Pa(X_i)$, a conditional probability table is specified as $p(X_i|Pa(X_i))$.

Given that the structure is a Bayesian network over $S = \{X_1, X_2, \dots, X_n\}$, the joint probability distribution for S is as follows:

$$p(S) = \prod_{i=1}^{n} p(X_i|Pa(X_i)) \quad (1)$$

where $Pa(X_i)$ denotes the parents of node X_i in S, and the parent node specifies that the arc is pointing from $Pa(X_i)$ to X_i. To be more exact, we can define the set of parents $\{Pa(X_1), \dots, Pa(X_n)\}$ corresponding to the Bayesian network parents for the variable set $\{X_1, X_2, \dots, X_N\}$. To infer the probability X_i (the variable of concern) that is in a certain state (e.g., 1) given observations of the remaining variables we compute the following:

$$p(X_i|e_1, \dots, e_k) = \frac{p(X_i, e_1, \dots e_k)}{\sum_{X_i} p(X_i, e_1, \dots, e_k)} \tag{2}$$

where $e_1, e_2, \dots, e_k$ are the new findings (updated states) of the other variables. To sum up, the joint probabilities can be computed from the Bayesian network given the set of network structure, S, and a set of conditional dependencies, P. To illustrate this, take note of the example giving by Figure 2-2. The conditional probabilities define the dependencies in the directed acyclic graph; $P(X_1)$, $P(X_2)$, $P(X_3|X_1, X_2)$, $P(X_4|X_3)$, $P(X_5|X_3)$, $P(X_6|X_4)$ and $P(X_7|X_4, X_5, X_6)$ are given either or are determined through some method. The probability of X_1 given observations of all other variables can be obtained as (in this case, $n = 7$):

$$\begin{aligned} p(X_1|X_2, \dots, X_7) &= \frac{P(X_1, \dots, X_7)}{P(X_2, \dots, X_7)} \\ &= \frac{P(X_1)P(X_2)P(X_3|X_1, X_2)P(X_4|X_3)P(X_5|X_3)P(X_6|X_4)P(X_7|X_4, X_5, X_6)}{\sum_{X_1} P(X_1)P(X_2)P(X_3|X_1, X_2)P(X_4|X_3)P(X_5|X_3)P(X_6|X_4)P(X_7|X_4, X_5, X_6)} \end{aligned} \tag{3}$$

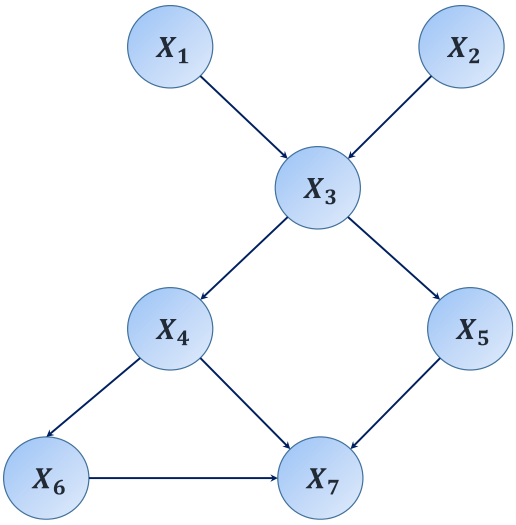


Figure 2-2 An example of a directed acyclic graph.

The induction of a Bayesian network from observations should adequately represent real observations, and the inferences should match the actual expected behavior of the system. Assessment metrics are required to measure how well a Bayesian network fits the data. The scoring metric returns a score reflecting the goodness-of-fit between the structure and the data. Cooper and Herskovit [164] assumed a uniform distribution of the network parameter and derived a uniform prior score metric, which is known as the K2 algorithm. Minimum Description Length (MDL) [165], [166] was proposed by Rissanen and Bouckaert to measure the quality of a network structure. Heckerman et al. proposed the Dirichlet distribution and derived the BDe metric [154].

CHAPTER 3. METHODOLOGY

A key objective of this research is to investigate the use of embodied interaction and its possible supporting role in solving spatial navigation problems while considering limitations of the user, both cognitive (human attentional resources) and physical (physical reach, biomechanical constraints of the body). Considering that a user's attention is a limited resource, generating a reasonable estimation for their level of attention (high or low) may help determine the most beneficial communication modality to leverage at this instance. In a scenario where a user interacts with a system via a specific modality of interaction and then the system delivers feedback to the user according to the activity at hand, a certain level of attention is expected to be allocated by the user in order to effectively complete the task. This feedback is the system's subsequent response – through a communication channel – to a command or function necessary to complete a task evoked by a user. To reach this goal, a Bayesian network (BN) is constructed to model the user's focus of attention in a given task. The representative Bayesian network can be obtained by: (1) the operators who are highly familiar with the task concerned (e.g., radiologist, intelligence analysts, and air traffic controllers), (2) adopting a genetic programming paradigm whereby the network evolves automatically as a result of iteratively varying genetic operations, (3) or a combination of both (1) and (2). In (1), experts or operators familiar with the task will construct the

Bayesian network manually. For the case (2), an evolutionary inspired approach for automatic network generation is employed. Once the best configuration of the Bayesian network is found, evidence is gathered to infer the user's state of attention. The evidence is comprised of values measured from various sensor outputs (or pre-processed cues, e.g., utterances, hand gestures and torso orientations) and task performance. Along with the inferred information concerning attention, the development of metrics – based on utility theory – to assess and derive suitable modalities for both interaction and feedback is the overarching goal of this dissertation.

A key feature of this work is dynamically inferring a user's level of attention in a non-intrusive fashion. This is done through the design of Bayesian networks – defined here as Bayesian Attentional Networks (BANs) – as well as their topology structure and parameters. Such models are designed to help infer the operator attentional levels during task performance.

This research is to propose a method of support for a user in solving spatial navigational problems through the use of natural and intuitive interactions. This involves the analytical determination of the most suitable combination of control and feedback that will eventually lead to a reduction in cognitive burden on the user; this would enhance his/her performance, and this in turn, may lead to better decision-making in complex settings.

This chapter will define the scope of our research problem, the problem statement as well as the model formulation to solve the proposed problem. This chapter begins with a description of the system architecture, moving on to an introduction of Bayesian networks for attention assessment, then to utility functions that express the cognitive, physical, and the technical benefits and costs of hybrid human-machine systems.

3.1 System Architecture

The architecture of the proposed framework is illustrated in Figure 3-1. Stages from A to F are shown in the Figure. (A) The task is geared towards the user solving a spatial navigation task, the Traveling Salesman Problem (described in Chapter 3.2). A navigational task is performed by the user employing gestures, speech, feet, or body stance in order to interact with a computing device. (B) Operators (users) complete the assigned task by evoking a sequence of commands using different modalities. While commands are evoked by the user, streams of signals (referred as observations or evidences) are collected by sensors. (C) The specific sensed signals include ambient acoustics detected by microphones, torso orientation and hand gesture detected and recognized via optical sensors (i.e., Kinect), and body stance configurations measured from pressure and weight sensors (dance pad controller, or Wii Balance Board). As the user completes a command, a feedback message is rendered to the user –using sound or visual cues – providing a performance metric (e.g. in the case of a navigation task, overall traveled distance. Armed with this information, the subject can better estimate the potential solutions leading to overall better performance metrics (e.g. shorter distances traversed). (D) Given the constructed Bayesian networks defined in Chapter 3.4, a discrete probability distribution describing the level of attention is computed by updating values gathered from evidence nodes and considering the conditional dependencies of all such evidences (based on the observations). (E) Utility theory is used to evaluate the trade-off relationships between task performance and the user's utility when feedback is provided. (F) Eventually, an enhanced interaction and feedback modality is assigned, for the purpose of maximizing operator's performance.

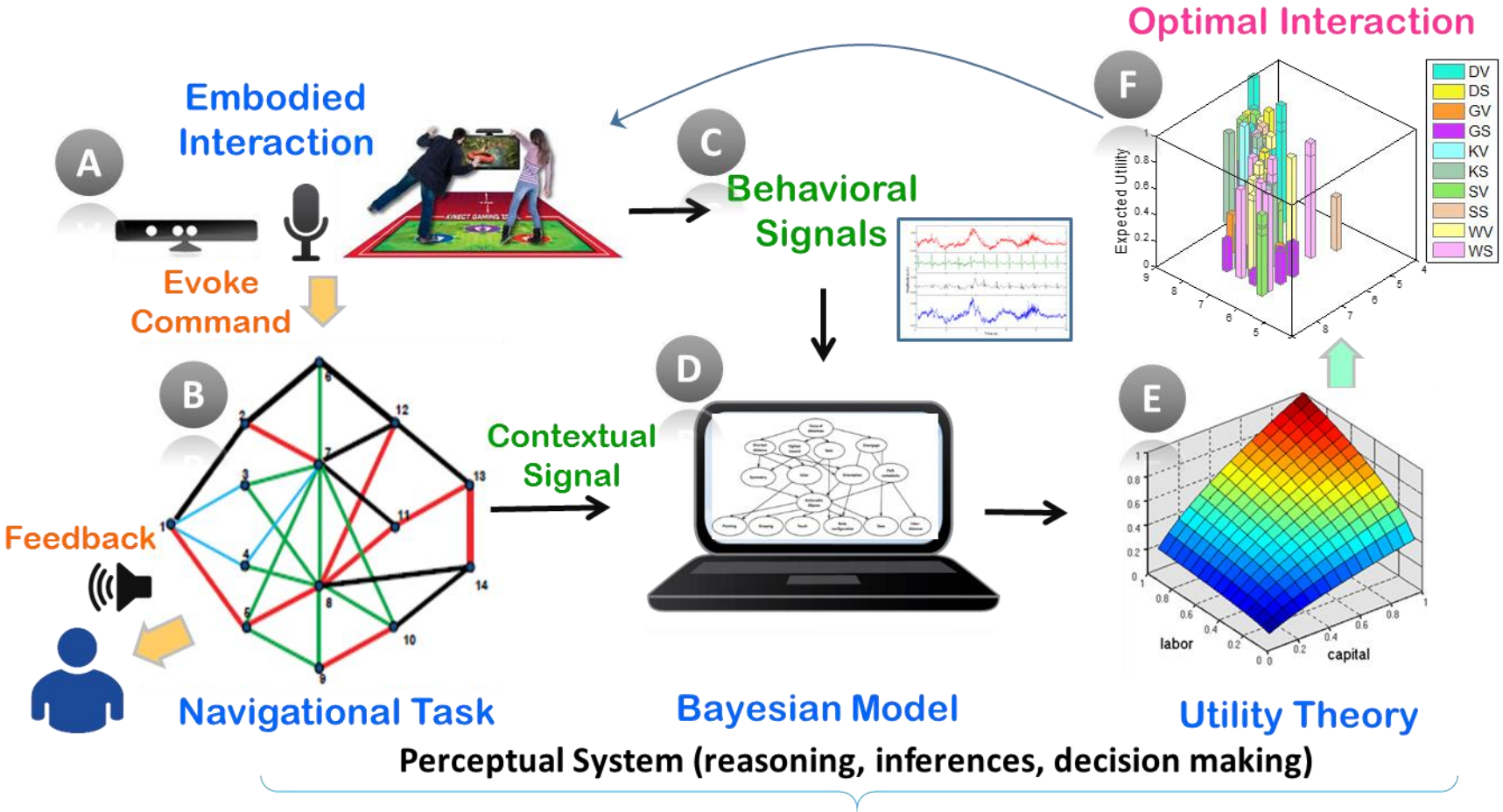


Figure 3-1 Framework for the study of embodied interaction through navigational experimentation, computational modeling, and decision-making.

3.2 Spatial Navigational Task

Spatial navigation refers to the ability of a user to navigate between focusable elements, such as hyperlinks and form controls, within a structured document or user interface according to the spatial location of such elements [167]. In this dissertation, the class of spatial navigation problems include traversing graphs, and solving those problems involve decision making and high attention associated with distributed networks. These spatial navigational tasks are time sensitive since the task is defined as a problem that must be solved as a function of time; this could either entail a time-dependent decrease in score or a limit to the absolute time allotted for the performance of the task. The users finishing the task faster will obtain higher benefits/rewards associated with their accelerated performance. More specifically concerning the type of spatial navigation task, we will focus on two navigational problems: (1) Traveling salesman problem (TSP), and (2) Cyber-physical navigational problems.

3.2.1 Traveling salesman problem

The traveling salesman problem (TSP) consists of having a group of cities (separated by physical distances between each pair); and the task requires a salesman to visit all the different cities (*N* cities) using the shortest trajectory without visiting any city more than once. This problem is simply described to and easily understood by uninitiated users, but requires high attentional focus to attain a near optimal solution (this is not trivial, especially for $N > 10$). Although computational methods have been suggested to offer solutions to this problem, there is no proof any such an algorithm has solved this problem optimally for a general number of cities. Studies [168], [169] indicate that people (and

animals) can obtain solutions that are near-optimal to TSP versions generated by computers. However, there is a significant variation in the strategies adopted by each individual. The generation of near-optimal solutions by various means is the reason why it is of paramount importance to investigate how humans solve this problem and what factors affect their solutions. For example, it was found that symmetry of the city layout and other aesthetic factors have an effect on the optimality of the solutions given by each individual [170]. In this dissertation, the TSP layout will follow the Symmetric [171] with Rewards setup [172], in which the distances between two cities are exactly the same in each direction; and there are prizes (rewards whose values decline over time) assigned to the cities (see Figure 3-2). In Figure 3-2, the distance between two cities is placed between the edges. The exponential decay, presented beside the circles (representing each city) expresses the change of reward assigned to the city as a function of time. In this context, the TSP is time sensitive; this is expressed as the faster the user visit the cities, the higher total rewards he can obtain. The goal is to find a path that minimizes the total distance while maximizing the total reward collected, subject to an overall limit on the total length of the path. The TSP is designed as a directed graph with a reward value, π_v, allocated at each vertex, v; the total reward function of visiting all vertices $v = 1 \dots N$ is expressed as:

$$\sum_{v=1}^{N} \pi_v \gamma(t_v) \tag{4}$$

where t_v is the time that has passed from the moment that the user starts solving the problem, $\gamma(t_v) = e^{-t_v/T_m}$ is defined as an exponential decreasing function and T_m is the maximum time allotted to solve the whole problem.

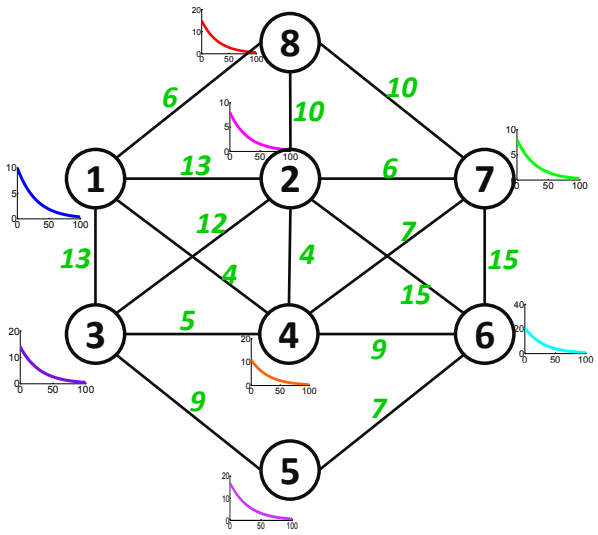


Figure 3-2 A 8-city TSP with reward at each city.

3.2.2 Cyber-Physical Navigational Problems

Cyber-physical network is a system of collaborating computational elements each of which controls physical entities such as: servers, robotics, computational engines, or a power grid. In this dissertation, a cyber-physical network represents the map of several United State Air Force Bases within the central United States (see Figure 3-3). The size of the nodes refers to the latency (how much time it takes for a packet of data to get from one designated point to another) of the base. The larger circle size means that the node is more congested and a data packet requires more time to transmit to the next node. In this problem, the goal is to transmit a data packet within the network through less congested nodes.

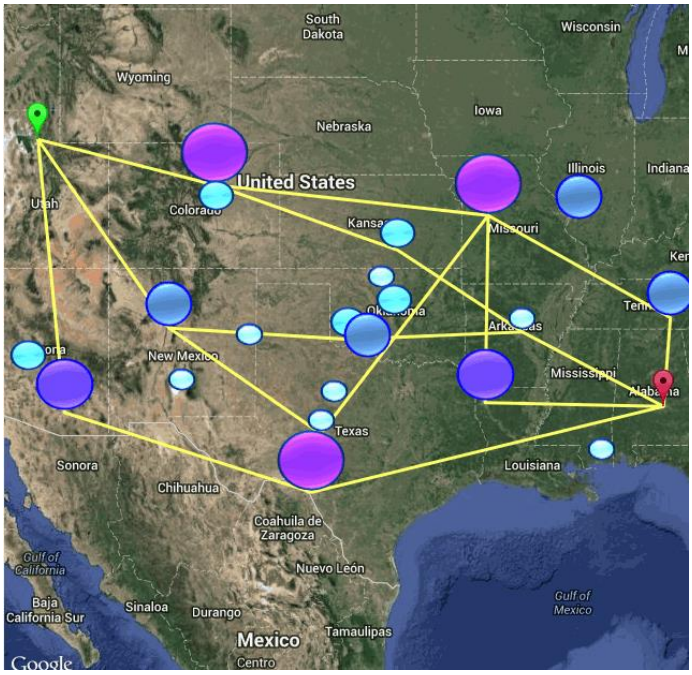


Figure 3-3 A cyber-physical network: network latency of US Air Force Bases.

3.3 Using Embodied Interaction for a Spatial Navigational Task

To complete the assigned navigational task, the user must use different interaction modalities to navigate the network. The detected command represents the next node which the user desires to visit. Various sensors were used in conjunction with the SDK to obtain an easy method for tracking human motion, including: the Kinect camera, microphone, and dance pad. The following five interaction modalities were considered (see Figure 3-4):

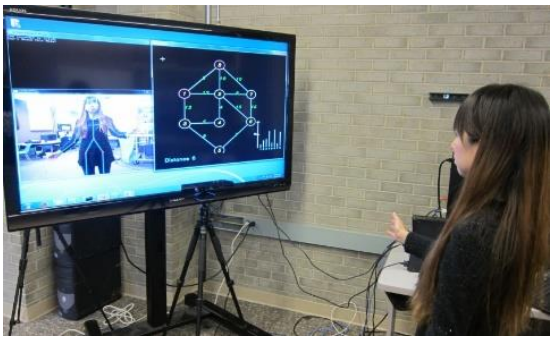
(a) Gross gesture movements: Using the arms and hands (e.g., wiping the hand from the center to right for the command “right”). The user’s arm movement is tracked by Kinect camera (see Figure 3-5).

(b) Fine gesture configurations: Using static hand poses (e.g., different figure configurations for different nodes). This gesture was detected by a data glove that user is wearing.

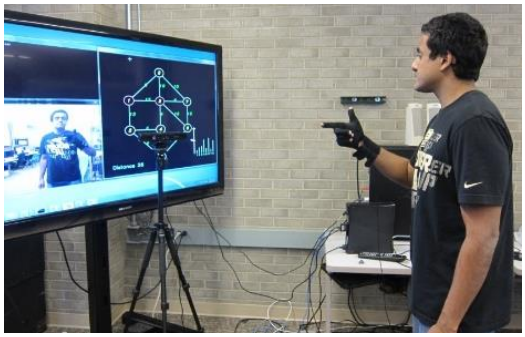
(c) Speech: Using spoken commands (e.g., “move left”). Audio was detected by a microphone.

(d) Foot gestures: Stepping over specific regions (e.g. jump right for “right”). A dance pad was used to detect the steps.

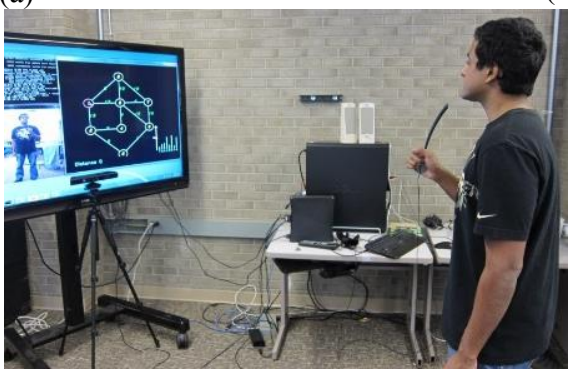
(e) Body stance: Changing the body balance (e.g. bending forward for “up”). A Nintendo Wii Balance Board was used to measure changes in the pressure.



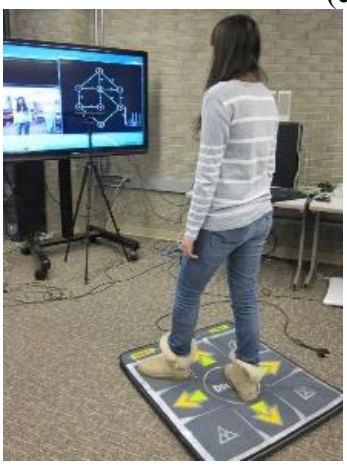
(a)



(b)



(c)



(d)

(e)

Figure 3-4 Five modalities used in the experiment. (a) gross gestures (Kinect) (b) fine gestures (glove) (c) speech (d) feet on dance pad (e) body stance on Wii balance board

The Kinect sensor can deliver a stream of images where body parts are tracked using a "skeleton" model [173]. This model can approximate the 3D coordinates of major joints in the skeleton (see Figure 3-5). This information was also used to build the Bayesian model that will be explained later.

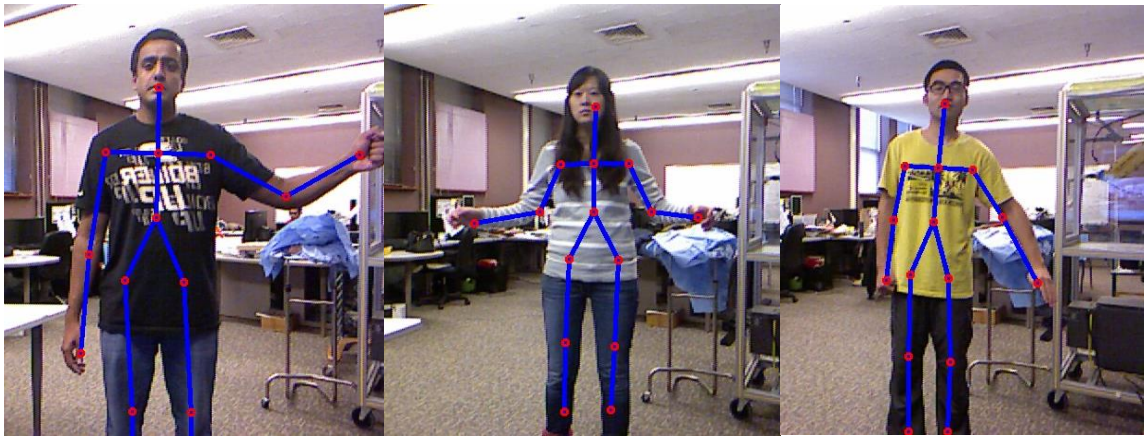


Figure 3-5 Kinect skeleton of a user.

3.4 Bayesian Attentional Network

In this section, a Bayesian model is used to represent the operator's attentional levels while solving spatial navigation problems. The model captures key cognitive processes characteristic to strategies used by the operators to solve decision-making problems; postures and actions during the decision making are used to thereby assess a user's level of attention. Figure 3-6 shows the system architecture of the Bayesian Attentional Network (BAN) framework. It is used to infer the user's level of attention based on the probability distribution of the query variable, attention (output), given values from evidence nodes, observations of physical actions and contextual information (inputs). Note that having contextual information (e.g., accuracy, task completion time) as part of

the BAN framework allows to model constructs above and beyond “physical engagement”. For example, it has been reported in previous research that “speed” and “accuracy” are directly correlated to operator’s attention [171][172].

In order to develop probabilistic models used to infer level of attention, a systematic approach is developed that integrates the operator’s knowledge with an automatic learning process. The enhanced BAN is further used to infer the probability of attention in different interaction scenarios. The representative Bayesian network, describing the operator’s attentional behavior, is obtained by: (1) the operators who are highly familiar with the task at hand [176] (e.g., radiologist, intelligence analysts, air traffic controllers); or by (2) adopting a genetic programming paradigm whereby the network evolves automatically as a result of iterative genetic operations towards an “incumbent solution”. An incumbent solution is a solution that is the best feasible solution known so far (not necessary “optimal”) for which all discrete variables can have discrete values [177].

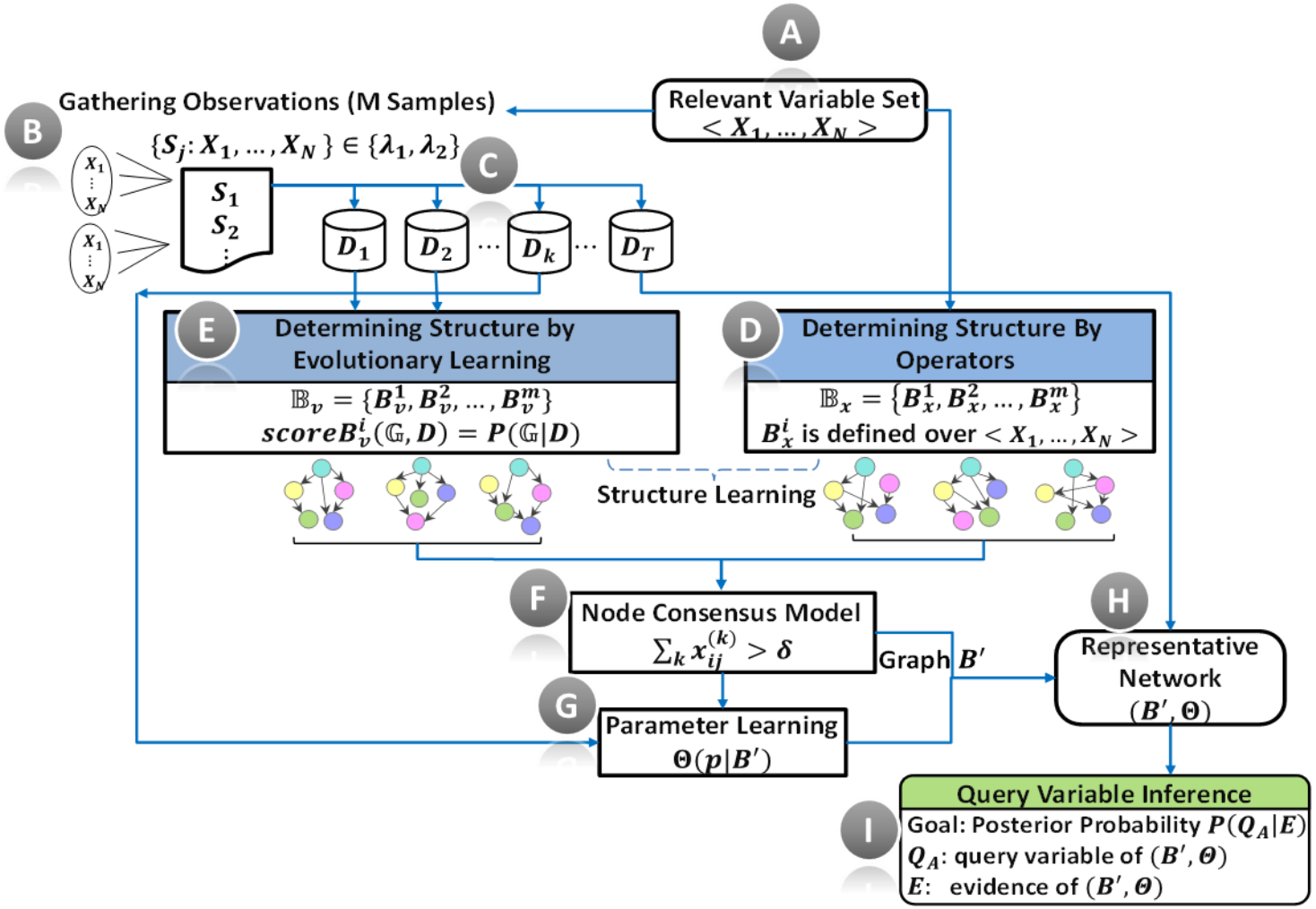


Figure 3-6 System architecture representing construction of the BAN.

The components from A to I within the framework are shown in the Figure 3-6. The structure of the BAN is defined as an assignment over N variables, $< X_1, X_2, \dots, X_N >$, each of which takes a binary value in a finite domain $\{\theta_1, \theta_2\}$ (label A). The description of a BAN (represented as B in Figure 3-6) consists of the directed acyclic graph, $\mathbb{G}$, which includes directed edges between variables and associated parameter vectors, Θ, that specify the associated conditional dependencies. The construction of the BAN consists of the following two steps: First, the set of N relevant variables are chosen to describe the problem domain. Secondly, the variables of interest are identified by knowledge engineering, domain experts or the system's operators.

In this dissertation, the variables mentioned above include observations of the user while they are solving the TSP problem using embodied interaction. Let us define a variable X_i (where $i = 1, .., k, .., N$) such that its value, $\lambda_i = f(X_i)$, is a Boolean. Also let X_1 be the query variable (level of attention). Attention can be discretized into states $\{0,1\}$, each representing "high attention", and "low attention", respectively. The sensors collect raw instances, S, about the user's physical behavior (body movements), and contextual information (e.g., accuracy, task completion time) during the experiment (label B in Figure 3-6). Those raw instances are transformed into the states' value of the variables, $f(S) \rightarrow X_2, X_3, \cdots, X_N$ (these are referred to as evidence variables). For each variable its definitions, description, and corresponding state are listed in Table 1. The variables are also discretized into the states $\{0,1\}$. This simplifies the total number of possibilities and values associated with each node. This quantization can be done in an empirical manner based on subjective assessments completed by the operators. An observation is defined as a feature vector, $\Psi = \{\lambda_2, \dots, \lambda_N\}$, where the binary value $\lambda_k = f(X_k)$ corresponds to the

X_k evidence variable computed from the operator's evoked command. The feature vector only contains the variables whose states are observable, and therefore λ_1 is not included (since it is an inferred quantity). To build a number of BANs through evolutionary learning, the overall observations were then randomly divided into an equal number of datasets, $(D_1, D_2, \dots, D_T)$, that were used in different stages while building the enhanced BAN (label C in Figure 3-6). Each dataset D_l is constituted by a number of feature vectors $\boldsymbol{\Psi} \in \mathbb{R}^M$, in other words, $D_l \in \mathbb{R}^{M,N-1}$, where M is the number of observations assigned to D_i. Some of the datasets were used to build the topology of the BAN based on an evolutionary learning algorithm, while the remaining datasets were used for parameter learning. In this case, a scoring metric was developed to determine goodness of fit between that dataset, D_l, and a given topology, $\mathbb{G}$. A different approach leverages the operators' knowledge (label D) and subjective assessment (label E) to design the BAN candidates. These approaches will be explained next.

Table 3-1 Definition of discrete states of each variable

Variable	Description	States
X_1	Level of Attention	{High Attention, Low Attention}
X_2	Torso Orientation	Detection of frontal torso {True, False}
X_3	Face Orientation	Detection of frontal face {True, False}
X_4	Hand Gesture	{Evoked, Not evoked}
X_5	Utterance	{Present, Not present}
X_6	Feet in Location	{Yes, No}
X_7	Inter-command Elapsed Time (t)	$\{\lvert t-\mu\rvert \leq \sigma, \lvert t-\mu\rvert > \sigma\}$ [1]
X_8	Error in Use	{Wrong command delivered, Correct command delivered}

[1] μ: mean of the inter-command elapsed time of all observation; σ: standard deviation of the inter-command elapsed time of all observation

3.4.1 Determining the BAN Structure by Operator's Knowledge

In operator-centered based modeling, the networks are constructed by operators who have domain knowledge, and who consider the systems' requirement and user-centric preferences. The procedure used by the operators for building the construction of networks is described in the algorithm below:

Algorithm 1: Constructing BAN by Operators

Input: A set of relevant variables, $< X_1, X_2, \dots, X_N >$, that describe the problem domain

Step 1. Start by placing the children nodes of the network (raw evidences) at the lower level. All these nodes are arranged in the same level

Step 2. Add the inferred node of the network at the top level, in our case: Attention.

Step 3. Assign a variable X_i with its description to each node in the network (descriptions are given in Table I).

Step 4. Add nodes in between the lowest level and the highest level, exhibiting a cause-effect relation. Work your way from the bottom to the top.

Step 4.1 For each node added, determine its connection between node X_i and the set of nodes already in the network.

Step 4.2 If a cycle exists, remove the last node.

Step 5. Return to Step 4 until all the nodes have been placed and all variables are assigned to nodes.

The idea of relying on the operator for the design of the BAN hinges on the operator having experience with the specific domain, effective problem solving within that domain, as well as being highly familiar with the interaction process itself. See examples of operator-based BANs in Figure 4-4 (a) – (e) in Chapter 4.1.1.

3.4.2 Determining the BAN Structure through Evolutionary Learning

In this section, evolutionary-based modeling was used for the construction of our Bayesian network. This method is found upon concepts within of Genetic Programming (GP), where the dependencies between nodes are inducted following operations from GP's. Assume that a graph, $\mathbb{G}$, consists of N nodes, where v_i indicates the i-th node. An arc $x_{ij} = (v_i, v_j)$ takes on binary values and equals one if it is directed from v_i to v_j or zero if it is not directed. The directed acyclic graph is then represented as a bit string, $x_{12}\ x_{13} \dots x_{2k} \dots x_{N-1,N}$ [178]. Figure 3-7 shows three examples of 3-node structure. Note that the bit string of this 3-node example is $x_{12}x_{13}x_{23}$. Thus the structure of 101 means that a 1 is assigned to x_{12} and x_{23} (nodes 1 and 2, and nodes 2 and 3, are connected, respectively) while x_{13} is assigned to a 0, communicating there is no connection between node 1 and 3.

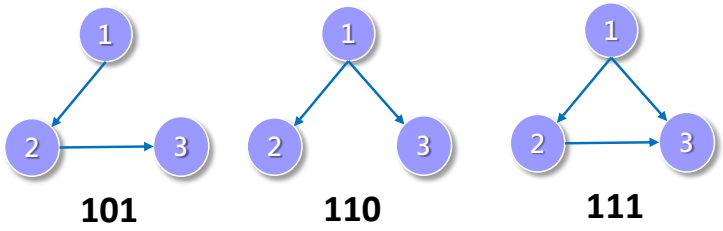


Figure 3-7 Node structure as bit representation $x_{12}x_{13}x_{23}$

In evolutionary-based modeling, first, an initial population was generated randomly. Then, selected individuals were used to generate a new generation. This was done through genetic crossover and mutation operators. Figure 3-8 shows these steps for a two BAN system:

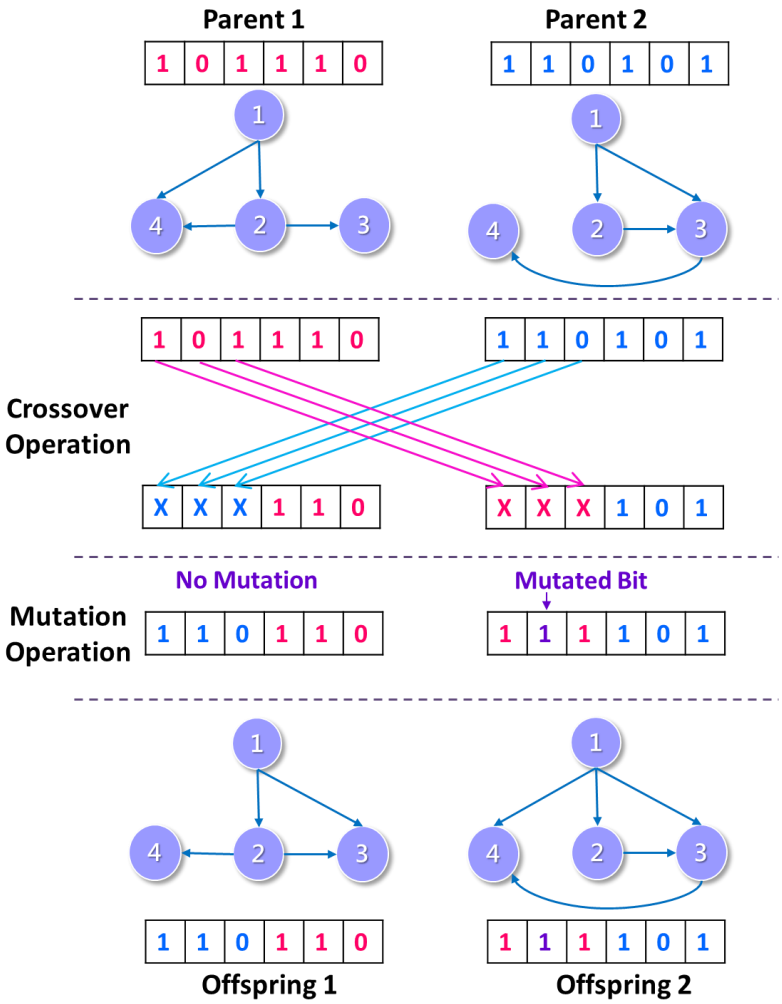


Figure 3-8 Example of crossover and mutation operations

The individuals remaining (each individual being a single Bayesian network) are those which outperform their antecedents in terms of a given performance function. This performance function acts as a cost function and is used to produce models of better fit in future generations. The fitness of the individual is assessed using a scoring measure in Eq. (5), which is the probability of observing the dataset, D_l, by an individual in each population [179]:

$$score(D_l, \mathbb{G}_H) = P(D_l|\mathbb{G}_H) = \sum_{i}^{2^M} P(d_i|\mathbb{G}_H) \tag{5}$$

$$P(d_i|G_H) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \frac{\Gamma(N_{ij})}{\Gamma(N_{ij} + M_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(a_{ijk} + s_{ijk})}{\Gamma(a_{ijk})} \tag{6}$$

Where within the above $\mathbb{G}_H = (V \cup H, E)$ represents the disjoint sets of observable variables ($V = \{X_2, \dots, X_8\}$) and the latent variable (level of attention, $H = \{X_1\}$), with edges E (between pairs of variables). A number of observation tables can be generated by concatenating the original table, D_l, with a new column, $c_i \in \mathbb{R}^{M,1}$ each time. More formally, $B_i = D_l \cup c_i$, $i = 1 \dots 2^M$, $B_i \in \mathbb{R}^{M,N}$. In this work, a user's level of attention (H) is inferred indirectly. The scoring metric $P(d_i|G_H, \Theta)$ in Eq. (6) is maximized through the Expectation-Maximization (EM) algorithm [179]. Two steps are needed iteratively from the current Θ to update to the next $\Theta^{(t)}$ at each iteration; these steps are as follows:

1. E-step of expectation: compute the probability expectation $q(\Theta^{(t)}|\Theta)$ of dataset d_i when current Θ is given:

$$q\left(\Theta^{(t)}\middle|\Theta\right) = E\left[P\left(d_i\middle|\Theta^{(t)}\right)|\Theta, d_i\right] \tag{7}$$

2. M-step of maximization: replace current Θ by:

$$\Theta = \arg\max_{\Theta^{(t)}} q\left(\Theta^{(t)}\middle|\Theta\right) \tag{8}$$

The notation used in Eq. (5) and (6) is as follows: M is the number of observations in D_l; r_i is the number of possible values of the discrete variable X_i (it is equal to 2 for Booleans); q_i is the total number of distinct possible values of the set of predecessors of

X_i; a_{ijk} is the parameter of a given Bayesian network with Dirichlet distribution; s_{ijk} is the number of samples in which X_i is equal to k and X_i's predecessors are equal to the j-th possible value; $N_{ij} = \sum_{k=1}^{r_i} a_{ijk}$ and $M_{ij} = \sum_{k=1}^{r_i} s_{ijk}$. As an example, take node 2 in the top left network in Figure 3-7. In this case, $i = 2$, $r_2 = 2$, and $q_2 = 2$ since the number of configurations for its predecessor, X_1, is 2 (0 or 1). Let us also suppose that the table below corresponds to the values of each of the variables in that network.

Table 3-2 Example of the values of X_i

X_1	X_2	X_3	X_4
1	1	0	1
1	0	0	1
1	1	0	0

Since s_{211} equals to the number of observations where $X_2 = 1$ and its predecessor is given as $X_1 = 1$, according to the table $s_{211} = 2$.

In Eq. (5), the computation of the scoring metric takes exponential time in terms of M. To tackle this problem, an efficient calculation [180] was carried out by computing $P(d_i|\mathbb{G}_H)$ for repetitive observations in the dataset only once, and multiplying the derived probability by the number of repetitions without affecting their statistical effect on the latent variable.

The overall procedure of evolutionary-based modeling for building our networks is described in the algorithm below:

Algorithm 2: Constructing BAN through the Evolutionary Learning Approach

Input:
    Table $\boldsymbol{D_l}$ – binary values of observable variables
    M – number of iterations; i – iteration index
Initialization: generate a set of feasible $\mathbb{G}_H$[2] solutions randomly
while score(D_l, $\mathbb{G}_H^{(i)^*}$) - score(D_l, $\mathbb{G}_H^{(i-1)^*}$) $\geq \epsilon$ **do**
    $\mathbb{G}_H^{(i)} \leftarrow crossOver\left(\mathbb{G}_H^{(i)}\right)$
    $\mathbb{G}_H^{(i)} \leftarrow mutation\left(\mathbb{G}_H^{(i)}, p_m\right)$ // p_m *as mutation probability*
    $\mathbb{G}_H^{(i)^*} \leftarrow eliteSelection\left(\mathbb{G}_H^{(i)}\right)$
    if $any\left(\mathbb{G}_H^{(i)}\right)$ is infeasible **then**
        update $\mathbb{G}_H^{(i)}$ // replace a infeasible solution by a new random one
    end if
    increment i
end while
Output: Incumbent DAG $\mathbb{G}_H^{(m)^*}$

[2] a feasible $\mathbb{G}_H$ defines as a graph without a vertex that is not an endpoint of any edge

Examples of the implementation of this algorithm may be seen in Figure 4-4 (f) – (j), in Chapter 4.1.1.

3.5 Consensus (Majority) Model

An enhanced graph structure is obtained from the candidate BANs previously found using operator-based modeling and the evolutionary approach. The procedure proposed and used is referred as the *Consensus (Majority) Model (CMM)*, which consists of iteratively deriving a graph agreed upon by a majority of the candidates. Therefore, we seek for the largest agreement as possible and not necessary consensus. Consensus is the optimal case, and it is a particular case of agreement among networks.

The Random Sample Consensus (RANSAC) Algorithm [181] is an iterative method to estimate parameters for a model from a set of observations. We adopted the basic concept

of RANSAC concerning the selection of the instance subset (in our case BANs) that can be best described by the model's parameters. Further, the candidate models are those which meet maximum agreement among the inliers. The CMM is used to obtain a graph that represents the maximum agreement among the majority of the candidate BANs. The enhanced network is derived iteratively by examining the existence (and popularity) of edges among the BAN candidates. Assume there are K BANs in the candidate set and for each, an adjacency matrix $\boldsymbol{A_k}$, with each element x_{ij} where $i, j \in \{1 \dots N\}$, is constructed to represent it. This means that an entry "1" assigned to x_{ij} means that nodes i and j are connected, and "0" otherwise.

The representative BAN starts from an initial empty graph in which nodes are not connected (an adjacency matrix, $\mathcal{A}$, with all entities equal to 0). Let us hypothesize that there is an edge between nodes v_i and v_j. Then we ask how many of the remaining graphs agree with this hypothesis. Thus the existence of an edge is decided by iteratively examining the consensus among the remaining graphs. There is a consensus about the existence of a specific edge, if and only if the numbers of graphs which have the same connectivity exceed some threshold. Figure 3-9 shows the resulting adjacency matrix of the representative BAN after applying the CMM to 10 candidate BANs. Each entry in each adjacency matrix included only binary values, and thus the values for entry (i, j) can be at most 10. For example, the top left value indicates that 10 BANs agreed that there is a link (cause-effect) between attention and torso orientation. This process is summarized in Algorithm 3.

$$\boldsymbol{\mathcal{A}} = \begin{bmatrix} 0 & 10 & 7 & 5 & 5 & 4 & 3 & 4 \\ 0 & 0 & 4 & 6 & 2 & 3 & 4 & 3 \\ 0 & 0 & 0 & 5 & 6 & 3 & 0 & 6 \\ 0 & 0 & 0 & 0 & 4 & 3 & 5 & 6 \\ 0 & 0 & 0 & 0 & 0 & 4 & 5 & 8 \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 0 & 0 & 2 & 0 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Figure 3-9 The adjacency matrix of the representative BAN for the 10 candidate BANs.

Algorithm 3: Consensus (Majority) Model

Input:
    $\boldsymbol{A_k}$ – matrices representing a set of graphs, each with order N
    K – the number of BANs
for all $i, j \leq N$ **do** // *given i,j as the source and destination indices of nodes* x
    $nCon \leftarrow \sum_k^K x_{ij}^{(k)}$
    if $nCon > K/2$ **then** // *majority is more than 50% agreement*
        $\boldsymbol{\mathcal{A}(i,j)} \leftarrow nCon$
    end if
end for
$\mathcal{G} \leftarrow$ Mat2Dag($\mathcal{A}$) // *convert the adjacency matrix to the directed graph*
$\mathcal{G} \coloneqq$ enhanced graph with majority consensus
Output: enhanced graph $\mathcal{G}$ with adjacency matrix $\boldsymbol{\mathcal{A}} = [x_{ij}]$

Once the CMM delivers the BAN that best represents all the models found previously, it can be used to assess attention levels based on the evidence computed from empirical experiments (see Chapter 4). Once the attention is computed from specific modalities of interaction, feedback channels and contextual information (which are represented through the evidence nodes), the next goal is to determine the utility of such combination of interaction and feedback forms. This goal is accomplished through a Utility-Directed Feedback Model, which is the focus in the next section.

3.6 Utility-Directed Feedback Model

Providing useful and effective feedback to the user has a direct effect on task performance. While this feedback can improve a user's performance, it can also cause distractions. Thus, the "cost" of this feedback is context-dependent, as well as user-dependent. It is necessary to adopt an analytical approach for the determination of benefits associated with an interaction and its weight compared to the cognitive costs when feedback is provided. In this dissertation, an approach is proposed based on the observations obtained from the navigation task, to analyze the costs and benefits of various performance metrics. The benefits and costs are functions with multiple dimensions that involve various performance measures [182]. These functions are later weighted by the attention levels (computed through the methods defined in chapter 3.4 and 3.5) and modulated by the interaction modalities used to give an enhanced feedback.

3.6.1 Cost and Benefit Metrics

Delivering improper feedback when the user exhibits a low level of attention may result in a decrease of task performance and completion, (due to the increased context-cognitive cost required for the user to understand the meaning of the feedback) and this is reflected by a certain associated cost. During low levels of attention periods, there are fewer cognitive resources available. Therefore it is desirable to make the best use of those few cognitive resources by providing the appropriate feedback. Alternatively, the benefit of using an appropriate feedback method matching the user attentional level and his/her modality of interaction has the potential of improving the overall task performance. To design the relation functions between benefits/costs and the task performance, four

performance metrics are used. Let us define B_i and C_i as the design benefit and cost associated with performance metric i ($i = 1,,..,4$):

1. Recognition ($i = 1$):

B_1 is the true hit rate for classifying the given modality; C_1 as the false positive rate of detecting the given modality. For example, if the control form is gesture-based, then the accuracy of recognizing the evoked gestures can be obtained using the computer vision algorithms [183]. The same algorithms can also provide false recognitions or false alarms values for the gesture interaction.

2. Time ($i = 2$):

B_2 is the saved time (time difference between maximum allotted task time and actual time spent); C_2 is the preparation time (the time that takes before a command is evoked). The saved time is all the time that the user "did not use" to find a solution for the TSP. If the operator uses all the allocated time for a trial, then this value is zero. The preparation time, instead, is the time that takes the user to move the body, arms and head to the desired configuration to trigger a navigation command. If Brian Computer Interfaces were used, the preparation time is considered to be significantly higher than the other modalities so far explored.

3. Quality of solution ($i = 3$):

B_3 is the reward obtained during the task (e.g., in the TSP rewards are given by visiting cities at specific times); C_3 is the difference between the actual distance traversed and the shortest distance (the optimal solution of the TSP). The rewards correspond to the dynamic computation of Eq. (4).

4. User's experience ($i = 4$):

B_4 is the user's subjective satisfaction about the modality used; C_4 is the subjective frustration experienced by the user. Both are obtained by questionnaire. The complete questionnaire used in this study is presented in Appendix A.

3.6.2 Expected Utility Function

The net value of performance metric i can be computed as the sum of benefits minus costs, $B_i - C_i$ [182]. The utility obtained by measuring the performance metric i is expressed as a linear function of both $B_i - C_i$:

$$U_i(B_i - C_i) = (B_i - C_i)/P_{i,max} \quad (9)$$

This difference is normalized by dividing through by $P_{i,max}$ which is the maximum level of performance metric i. Thus, the expected utility function $U(I_k, F_j)$ associated with interaction modality I_k and feedback modality F_j is given by:

$$U(I_k, F_j) = \sum_i \omega_i U_i(B_i - C_i) \quad (10)$$

where ω_i is the weighting factor assigned to performance metric i. (the importance that the decision maker assigns to that metric)

Our goal is to find the feedback and interaction modality (considering the user's level of attention and task performance) which yields the highest utility. High levels of attention contributes more to user utility than do low levels of attention. The level of attention is represented as a discrete probability distribution. Wherein, the greater the likelihood of high attentional level, the higher the utility obtained with these performance measures, as

well. Thus, by multiplying the probability of high attentional level by the expected utility (Eq. (10)), optimal interaction and feedback modalities can be determined. The most suitable modalities are those which maximize the expected utility function considering the probability distribution of level of attention (X_1) given the observed evidences $\boldsymbol{e}$, and performance metrics:

$$\underset{I_k,F_j}{\operatorname{argmax}} \sum_i \omega_i U_i (B_i - C_i)\, p(X_1 = \theta_1 | \boldsymbol{e}) \tag{11}$$

where the probability is inferred by the representative BAN with N variables, $< X_1, X_2, \dots, X_N >$, each of which takes a binary value within a finite domain $\{\theta_1, \theta_2\}$.

Once the expect utility is computed for each modality and feedback form, it is possible to tell what combinations of interaction and feedback lead to the highest performance metric. This is dependent on the task selected. In the next chapter, we will discuss two tasks for which these methods were applied: (a) the TSP, and (b) a Cyber-physical threat avoidance system.

These tasks were completed by student subjects, and performance metrics were captured during this interaction. The next section reports the main procedures and findings throughout the completion of these two tasks.

CHAPTER 4. EXPERIMENTAL RESULTS

Experiments were conducted to test the effects of embodied interaction on task performance and its dependency on the user's attention levels. Two real-world experiments were conducted with visual interaction systems designed in two stages. Case Study 1 involves a systematic characterization of the operator's physical interactions while solving a spatial navigational problem (i.e., Traveling Salesman Problem), where the operators are able to navigate the visual environment. The best combination of interaction modalities and feedback was determined for dynamic decision making scenario. Case Study 2 is designed for visualizing cyber-operations in which operators interact with datasets of cyber-physical visual information using embodied interactions in a series of time-sensitive tasks. Institutional review board (IRB) permission (Purdue IRB Protocol # 1308013871) was sought and obtained to conduct these experiments.

4.1 Case Study 1

In Case Study 1, the experimental setting is designed for the users to solve TSP-type navigation problems. We conducted the experiments and collected observations while users were solving the TSP problem under varying conditions. The observations collected were further used to build the representative BAN, and several metrics were used to assess user's performance. In this experiment, we addressed RQ1 by determining the

optimal combination of interaction and feedback modalities in TSP-type spatial navigational problems.

4.1.1 Design of Experiments

Twenty graduate and undergraduate students were recruited, including 13 males and 7 females, all 20 – 30 years old. The users were given instances of the TSP problem to solve. Each user was issued 20 different TSPs divided into 4 different scenarios (5 TSPs in each scenario). In each scenario, we use a letter acronym to represent the type of modality: D, feet movement as interaction modality on dance pad; G, gesture with glove; K, gesture recognized by Kinect; S, speech; W, body stance measured by Wii Balance Board. There were also 2 feedback modalities: V, visual; and S, speech. Likewise, acronyms with the first letter of a modality and feedback, respectively, denote a single modality/feedback condition (e.g., "DS" means feet on dance pad as control, and speech as feedback modalities). Table 4-1 presented the summary of collected trials for each scenario.

Table 4-1 Summary of collected trials for each scenario.

	DV	DS	GV	GS	KV
20 subjects (13 males, 7 females)	5 trials / subject	5 trials / subject	5 trials / subject	5 trials / subject	5 trials / subject
	KS	SV	SS	WV	WS
	5 trials / subject	5 trials / subject	5 trials / subject	5 trials / subject	5 trials / subject

In each scenario, the subjects had to adopt a different interaction and feedback modality, which were randomly assigned in advance. In the beginning of each scenario, the subject

was given a single training attempt for the purpose of allowing the subjects to be experienced and be familiar with the scenario and settings. Beyond this training, learning effect is not studied in this experiment. Each user acted as an "operator", since their domain knowledge for solving the TSP is as good as anyone else's domain knowledge. The five modalities adopted included: gross hand gestures (using mainly the arms), fine hand gestures (using fingers configurations), speech, feet gestures (on dance pad controller), and body stance (using a Wii balance board); see Figure 3-4. Each city within each TSP was randomly assigned a reward value which decreased exponentially over time in accordance with Eq. (4). A sequence representing the decreasing reward of a city over time is presented in Figure 4-1. Each "active" cell in the 3x4 grid represents one of the cities. Brighter colors indicate higher reward values while darker colors represent lower rewards. As can be seen, all active cells become darker since rewards are reduced with passing time. In the experiment, the reward value assigned to each node is displayed as a bar plot to avoid confusions; see Figure 4-3.

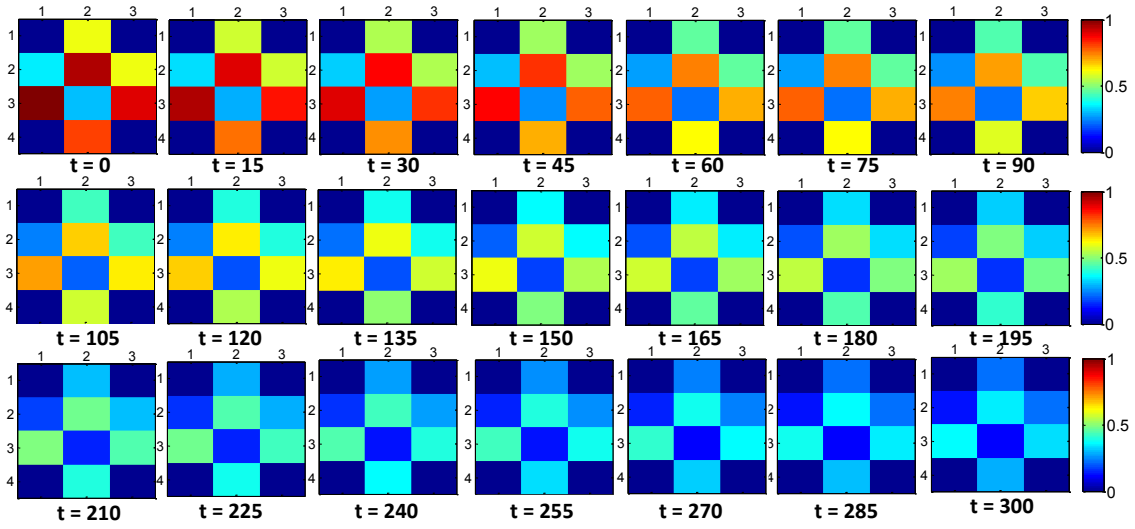


Figure 4-1 The reward of 8 cites discounted over time

Our experimental apparatus consisted of (1) a PC and a large 60" screen; (2) a speaker; (3) a Kinect sensor; (4) a data glove; (5) a microphone; (6) a dance pad; and (7) a balance board (see Figure 4-2). Those sensors were used to collect evidence including: torso and face orientations, hand gestures, utterance, body stance and elapsed time, which served as the raw observations (evidence) for the BANs. These constitute the user's input to the system. It is recognized that input devices differ in complexity of use. There may also be individual differences in operator-preferences in using the devices. Input complexity and individual-preferences is not studied in this experiment.

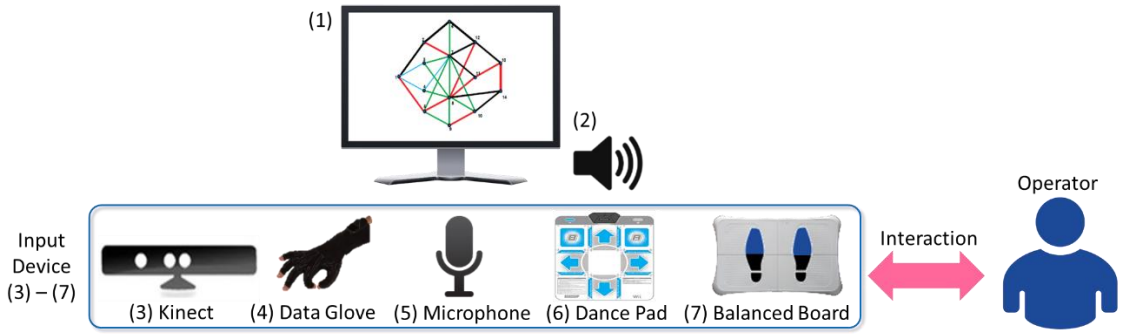


Figure 4-2 Experimental apparatus

The instances of the TSP problem presented included the layout of cities, labeled edges representing the distance between cities and the reward value assigned to each city (see Figure 3-2). As subjects move to the next city using one of the aforementioned interaction modalities, feedback is displayed or read back to them through a text-to-speech program, such as Microsoft SAM. The feedback information consisted of the overall travelled distance (see bottom left of Figure 4-3), and the rewards obtained (see bottom right of Figure 4-3). This information constitutes the output of the system. With

this information, the subjects were better equipped to estimate possible alternatives that would lead to shorter distances, i.e. better solutions.

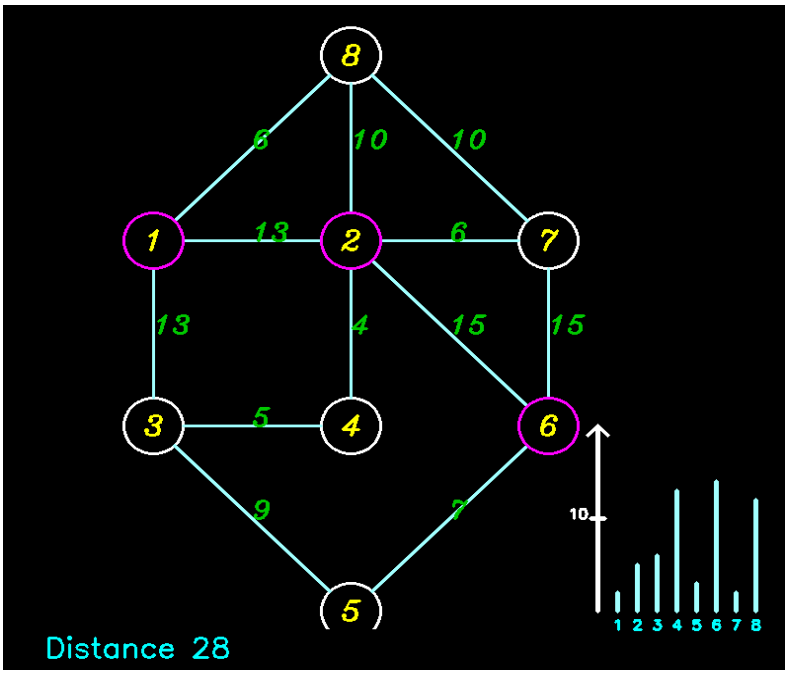
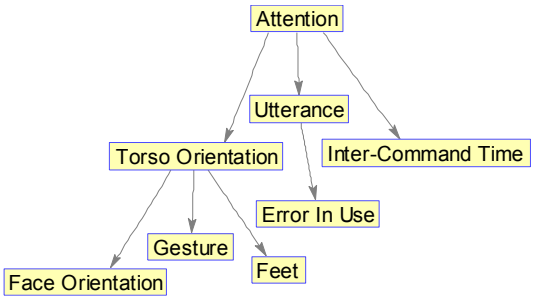


Figure 4-3 Visualized TSP displayed to the users

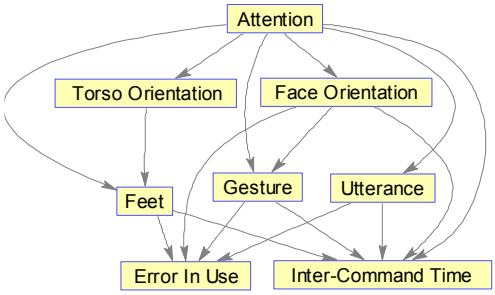
Subjects were assigned to random instances of the TSP to assess their problem solving. A trial is defined as a sequence of commands, $\{c_1, c_2, \dots, c_m\}$, required to solve an instance of the TSP. Each command c_j, results in an observation vector $\Psi = \{\lambda_2, \dots, \lambda_N\}$, defined previously in Chapter 3.4. In this manner, a total of 393 independent trials were collected. Each trial was designed to require 5 to 8 commands to complete the task. For example, an operator may complete a trial using 5 commands, we refer these commands as observations. From those all trials, 193 trials were used to create a training dataset of 1200 observations, and the remaining 200 trials resulted in 1670 observations.

4.1.2 Results: Bayesian Attentional Networks

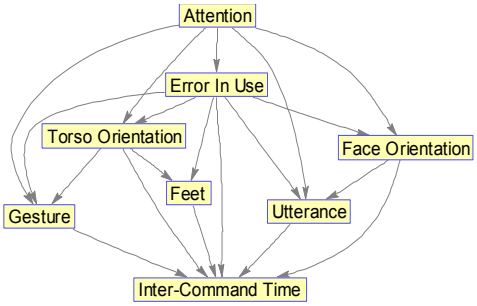
Five topologies were acquired using the evolutionary BAN approach from 40 observations in each dataset. Additionally another five BANs were obtained by operators. The parameters (conditional probability distribution for each node) that quantify relationships between connected nodes were computed using the Expectation-Maximization (EM) algorithm [179]. Figure 4-4 (a) – (j) shows the BANs constructed by 5 operators and learned through the evolutionary process.



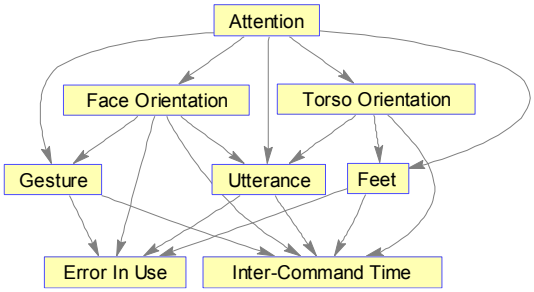
(a) Operator 1, $\text{score}(\mathrm{D}, \mathbb{G}_{\mathrm{H}}) = 0.65$



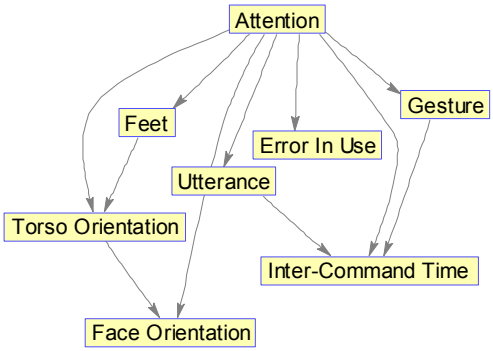
(b) Operator 5, $\text{score}(\mathrm{D}, \mathbb{G}_{\mathrm{H}}) = 0.69$



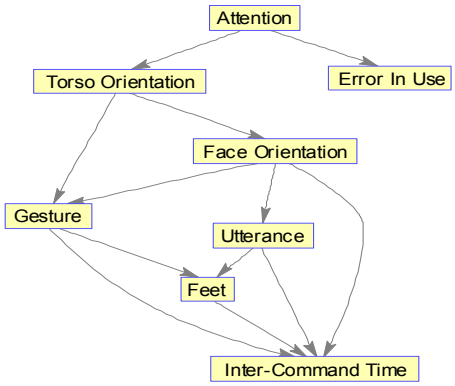
(c) Operator 3, $\text{score}(\mathrm{D}, \mathbb{G}_{\mathrm{H}}) = 0.60$



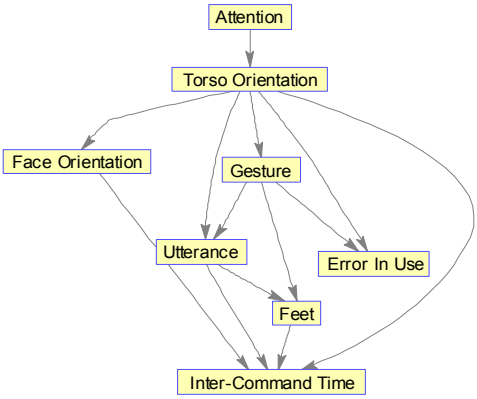
(d) Operator 4, $\text{score}(\mathrm{D}, \mathbb{G}_{\mathrm{H}}) = 0.68$



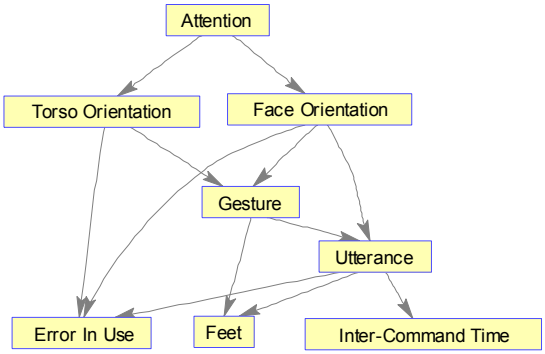
(e) Operator 5, score(D, $\mathbb{G}_H$) = 0.64



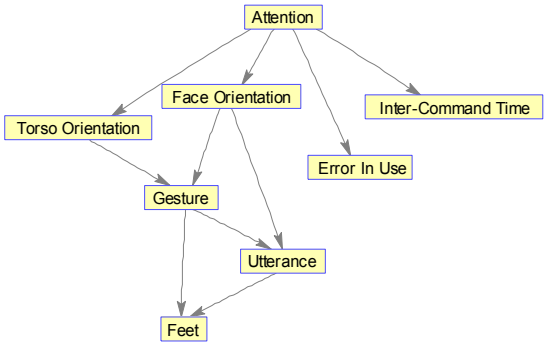
(f) score(D, $\mathbb{G}_H$) = 0.864



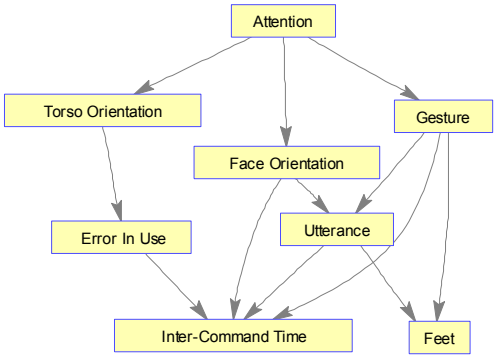
(g) score(D, $\mathbb{G}_H$) = 0.926



(h) score(D, $\mathbb{G}_H$) = 0.901



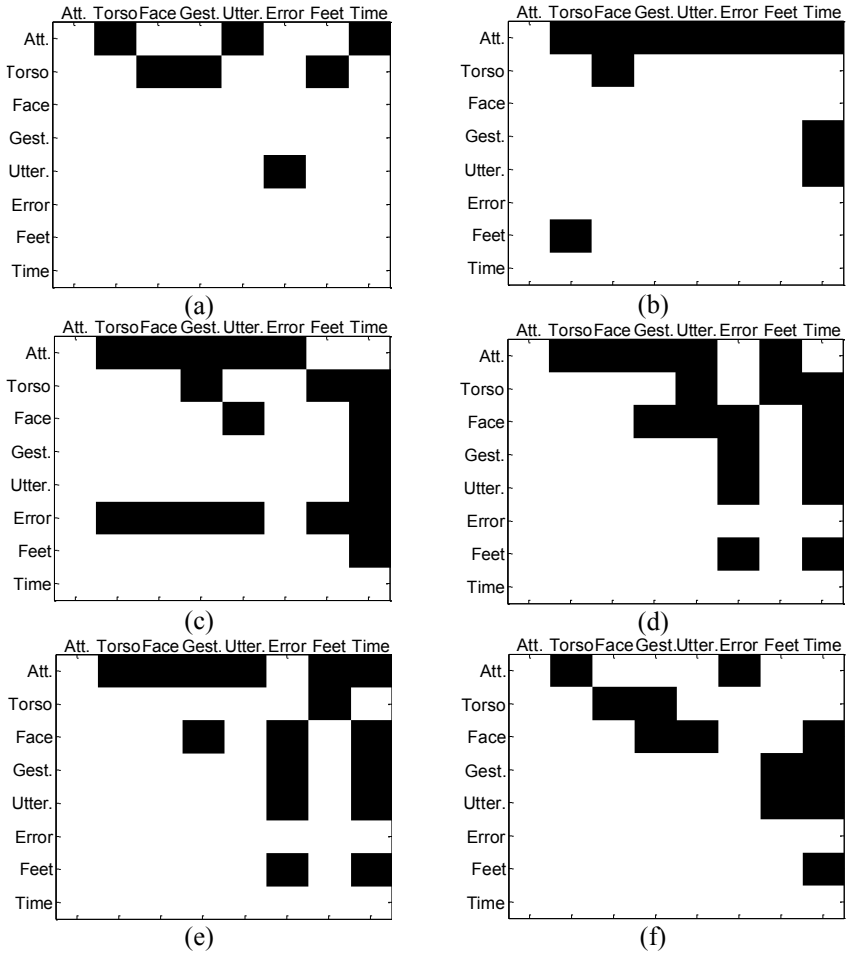
(i) score(D, $\mathbb{G}_H$) = 0.843



(j) $score(D, \mathbb{G}_H)$ = 1.000

Figure 4-4 Bayesian Attentional Network's structure obtained by (a) – (e) Operator based, (f) – (j) Evolutionary learning.

The adjacency matrix displayed for each BAN represents the links between nodes. In Figure 4-5 (a) – (j), a black cell in column *i* and row *j* is an entry of "1", meaning that node *i* and *j* are connected by a link, and a white cell is assigned to an entry of "0", denoting no connection between the nodes. The representative BAN determined by CMM method is shown in Figure 4-6 (a), and its adjacency matrix was shown in Figure 4-6 (b).



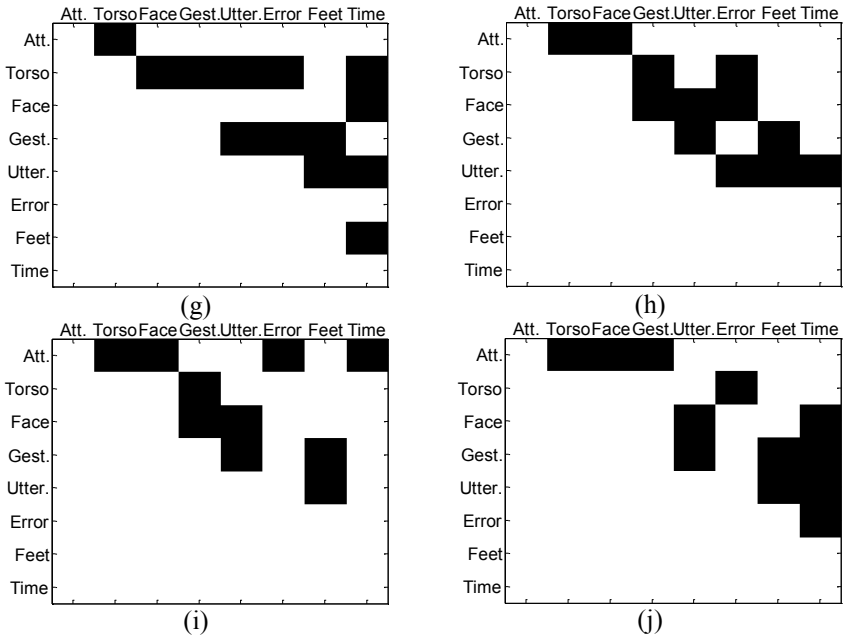
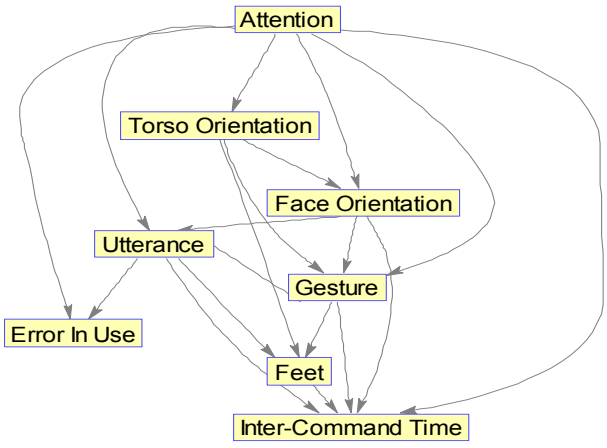


Figure 4-5 The adjacency matrix of BANs obtained by (a) – (e) Operator based, (f) – (j) Evolutionary learning



(a)

$$\mathcal{A} = \begin{bmatrix} 0 & 10 & 7 & 5 & 5 & 4 & 3 & 4 \\ 0 & 0 & 4 & 6 & 2 & 3 & 4 & 3 \\ 0 & 0 & 0 & 5 & 6 & 3 & 0 & 6 \\ 0 & 0 & 0 & 0 & 4 & 3 & 5 & 6 \\ 0 & 0 & 0 & 0 & 0 & 4 & 5 & 8 \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 0 & 0 & 2 & 0 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(b)

Figure 4-6 Representative BAN and its enhanced adjacency matrix

The resulting network obtained through the CMM method (Figure 4-6) displays how the level of attention affects the physical action as well as the task performance (elapsed time, and erratic commands). It can be observed that the torso orientation determines largely

the direction where the user is facing and the involvement of her feet movement in performing a task. The gesture and utterance were determined sufficiently by the orientation of the user's face (which in turn is a proxy of level of attention). The elapsed time varied among users depending on the time taken to evoke the necessary gestures or utterances.

Figure 4-7 shows the evolutionary learning process of five BANs for each generation. There are five evolutionary BANs generated through Algorithm 2, and the best (G_i^*) scores among the populations in each generation were plotted. Each curve presents the evolutionary process of a BAN.

This figure shows the convergence characteristics of the evolutionary learning approach; also shows the best scores among the populations within each generation. From Figure 4-7 can be learned that after 170 generations, the solution increased significantly (25.08% at most, and 9.77% at least) from their initial values.

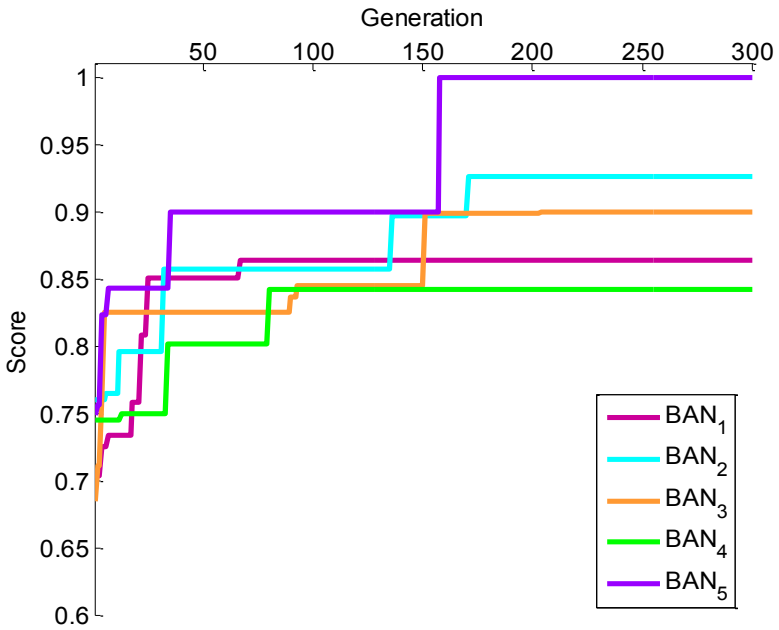


Figure 4-7 Convergence characteristics of 5 evolutionary BANs.

4.1.3 Enhanced Interaction Modality

The utility of the interaction used by the tested subjects was computed by following Eq. (9) – (11) based on the testing data. The testing data consists of 10 scenarios (5 interaction modalities and 2 feedback modalities) with 20 samples for each scenario. Post hoc power analysis for a significance level of 5%, is over .99 for that sample size. The testing data was randomly assigned to subsets, as detailed in Chapter 3.4. We use a letter acronym to represent the type of modality: D, feet movement as interaction modality on dance pad; G, gesture with glove; K, gesture recognized by Kinect; S, speech; W, body stance measured by Wii Balance Board. There were also 2 feedback modalities: V, visual; and S, speech. Likewise, acronyms with the first letter of a modality and feedback, respectively, denote a single modality/feedback condition (e.g., "DS" means feet on dance pad as control, and speech as feedback modalities). In order to show the optimal scenario, or alternatively the worst, to be significant the ANOVA (Analysis of variance) is conducted on each independent trial. Results of one-way ANOVA ($F(9,190)=58.75$, $p< .001$) indicated that there are statistical differences between the means of sample groups. Repeated Measure Analysis has been conducted and no significant changes in the interaction's utility over repeated trials ($p > .05$) was found. Figure 4-8 shows the boxplot of the expected utility for each trial within 10 different scenarios. The top of each box are the first and third quartiles, and the band inside the box represents the median. The ends of the whiskers associated with the boxes represent the minimum and maximum of the utilities.

To determine which specific sample groups differ, further post-hoc test was conducted. Figure 4-9 presents our findings from a Dunnett's test over all 10 combinations of

interaction-feedback modality pairings; we present these combinations as parings of the letter-representations introduced above, i.e. DV represents dance pad interaction coupled with visual feedback. We have found that the DV pairing displays the highest mean value for level of utility. The confidence intervals associated with the DS, KS, WV, and WS interaction-feedback pairings overlap with that of the DV sample case; we may not make any absolute assertions concerning these sample cases. All remaining pairings (GV, GS, KV, SV, SS) can be said to be significantly lower in utility than the DV pairing. Analyzing the mean values alone, the dance pad performs better than all other forms of interaction. It is evident from our data analysis that the glove-based (i.e., fine gestures) interaction modality performs significantly worse than all others; this suggests that the more gross gesture (dance pad, Kinect, and Wii Balanced board) interactions allow for better performance from the user. It is also evident that no clear statement may be made concerning which feedback modality has greater utility, as no case shows either S or V significantly outperforming the other. Note that these observations are limited to the sample groups in our experiment.

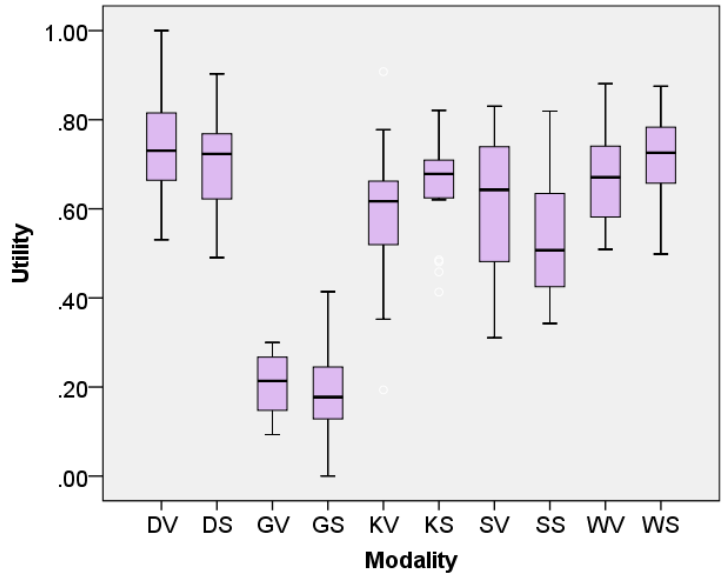


Figure 4-8 Boxplot of 10 interaction scenarios

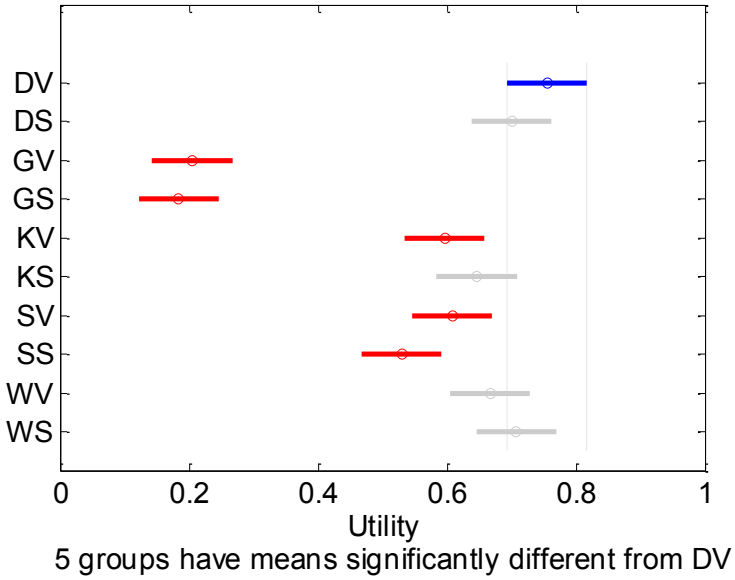


Figure 4-9 Group means comparison

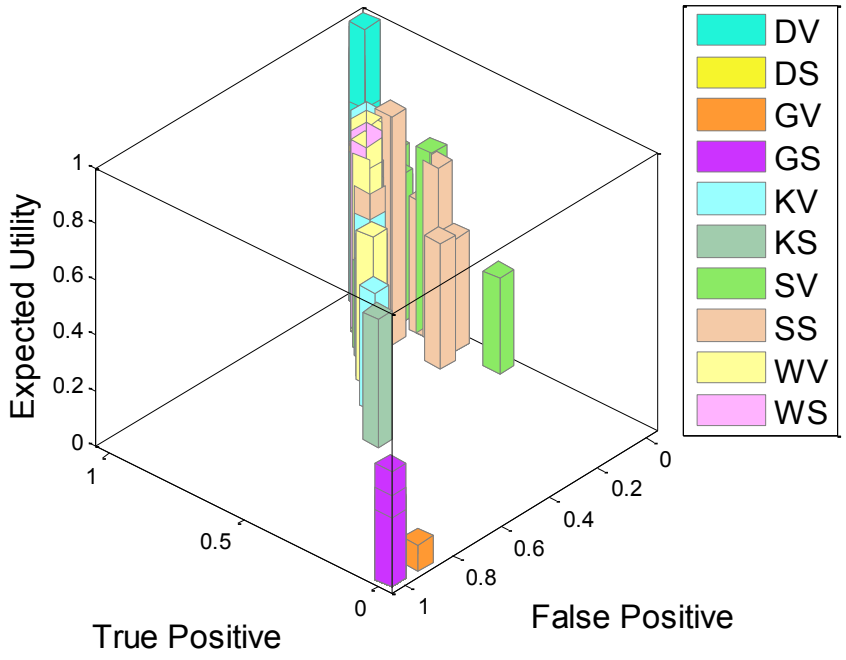
4.1.4 Task Performance of Interaction and Feedback Modality

Several metrics of task performance including recognition rate of interaction modality, total task completion time, preparation time, and solution quality are measured during the

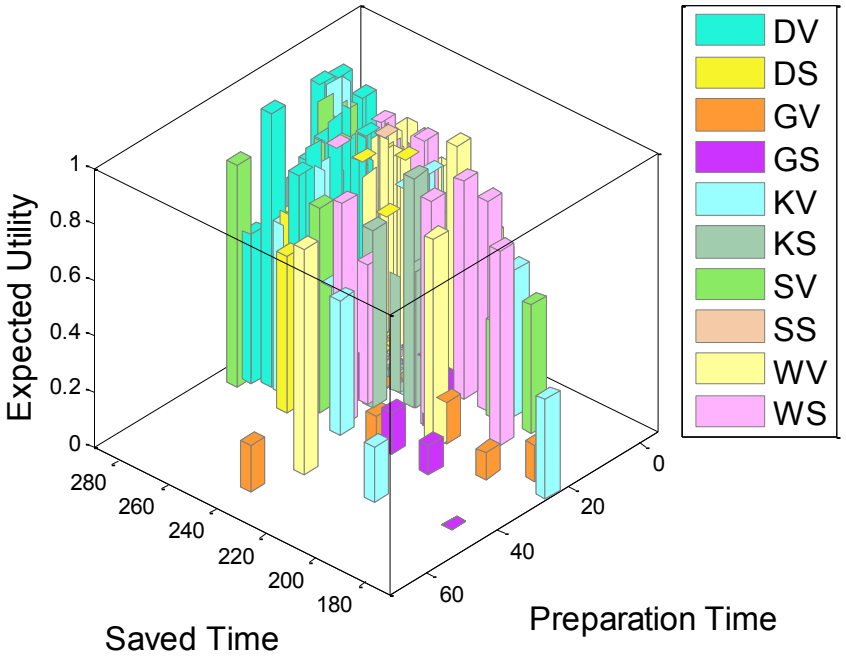
experiment and further compared. This was done to discover the relationship between interaction forms, expected utility and performance. Additionally, a satisfaction survey was administered after task completion. The relationship between the utility of interaction and performance metrics was determined through post experimental data analysis, see Figure 4-10.

The expected utility versus recognition rate is shown in Figure 4-10 (a). The DV modality (dance to interact and visual feedback) received the highest average true positive rate and lowest average false positive rate. The expected utility versus performance metric: B_2, saved time and C_2, preparation time is plotted in Figure 4-10 (b). It is shown that instances obtained from the DV modality resulted in higher expected utility as well as shorter completion time and preparation time.

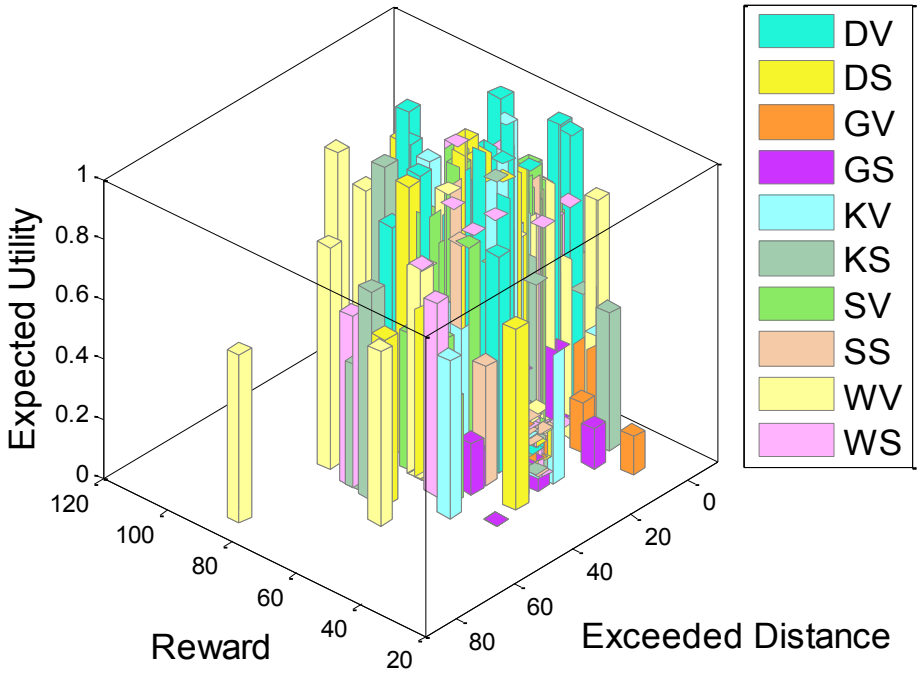
Bar plots displaying the expected utility versus: B_3, reward and C_3, exceeded distance (the difference between optimal and incurred distances) are presented in Figure 4-10 (c). The DV modality was also shown to deliver a better solution, i.e. a higher reward and shorter exceeded distance. As shown in the Figure 4-10 (d), the DV modality also received a higher user satisfaction score. Better task performance is associated both with higher attentional level and a higher interaction utility.


Expected Utility
1
0.8
0.6
0.4
0.2
0
1
0.5
0
True Positive
0
0.2
0.4
0.6
0.8
1
False Positive
DV
DS
GV
GS
KV
KS
SV
SS
WV
WS

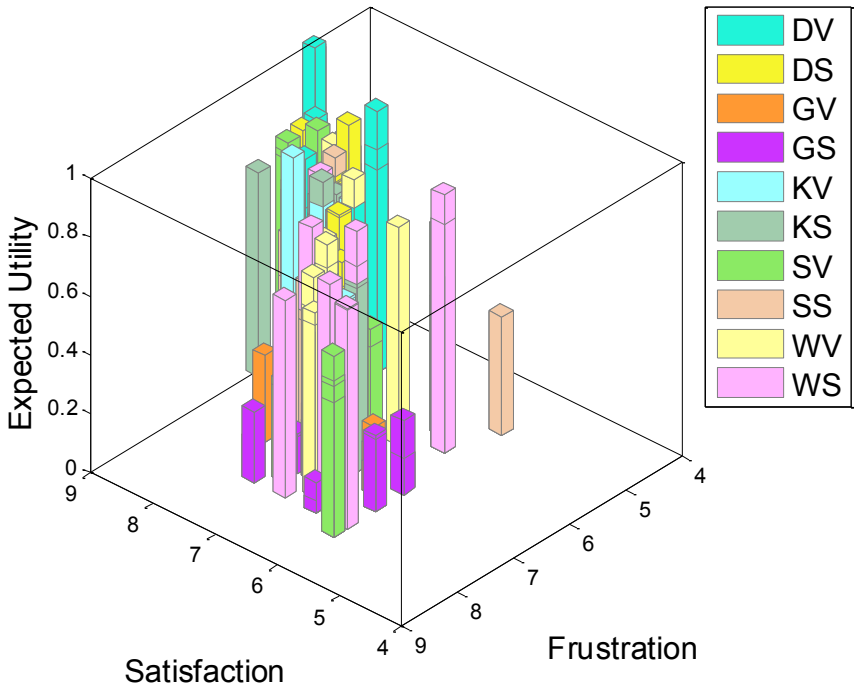
(a)


Expected Utility
1
0.8
0.6
0.4
0.2
0
280
260
240
220
200
180
Saved Time
0
20
40
60
Preparation Time
DV
DS
GV
GS
KV
KS
SV
SS
WV
WS

(b)



(c)



(d)

Figure 4-10 Expected utility vs. benefits and costs

4.2 Case Study 2

In Case Study 2, a cyber-physical network was displayed and designed to let users interact with cyber-physical scenarios. A multimodal system that allows the user to use embodied interaction (gross hand gestures, speech, or feet gestures simultaneously) was implemented for this purpose. The subjects were asked to perform a series of cyber-physical operations using the multimodal system or the keyboard system. The Graphical User Interfaces (GUI) was implemented through Qt, a cross-platform application framework that is widely used for developing application software GUI [1], with combinations of Google API. The design of the task is as follows:

1. A map of United State Air Force Bases located within the central United States, the relative size of the base representing its intra-network's congestion level (see Figure 4-11). The user's goal is to transmit a data packet in the network through the less congested nodes from the origin (green marker) to the destination (red marker). This is accomplished by selecting a path between nodes, within the network displayed. The operators are performing a time sensitive task since the goal is to minimize the total time spent in transmitting a data packet within the network.

[1] http://en.wikipedia.org/wiki/Qt_%28software%29

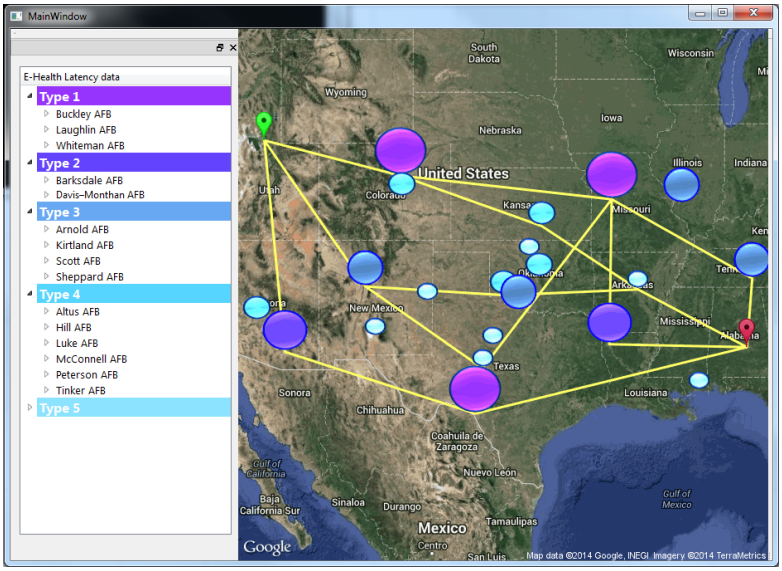


Figure 4-11 Cyber operation tasks: layer 1

2. To accomplish the previous task, the next step is to test whether or not the user can identify and select which direction has the largest amount of data packets to be transmitted (see Figure 4-12). The users have to browse through different categories of data packets (on-access scan, on-demand scan, web threat, mail threat, instruction detection scan, and vulnerability scan), which are represented by different colors, and select the linkage with largest packet size within each category. Data packet size is depicted by the thickness of the line.

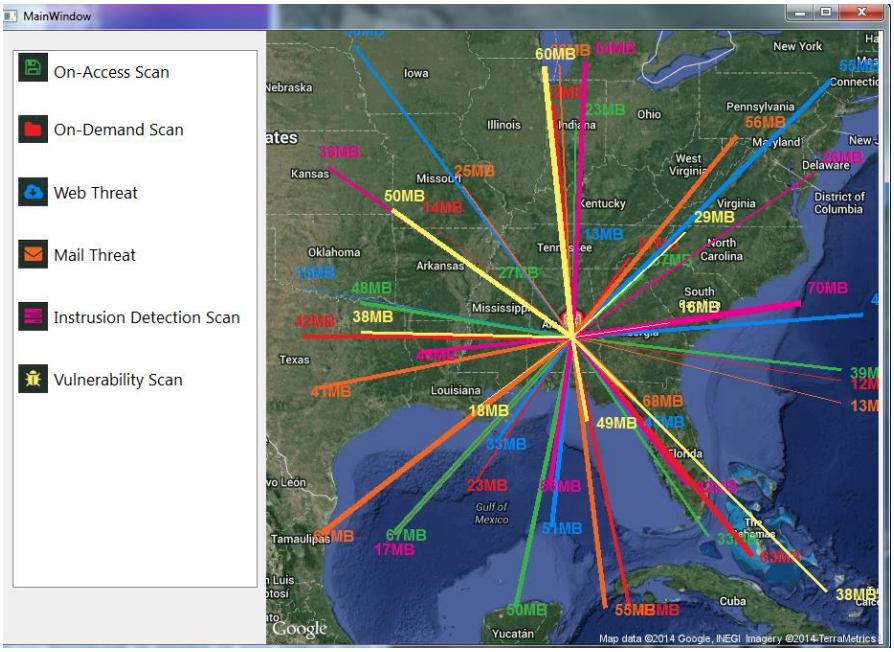


Figure 4-12 Cyber operation tasks: layer 2

3. Finally, the user needs to assess the location and name of the machines that are under threat. The users have to traverse through the network and then select those machines which are marked in red, i.e., at risk, using a modality of interaction; a representation of this task can be seen in Figure 4-13. Figure 4-14 shows the prototyped and implemented multimodal interface that enables users to interact with a cyber-physical system.

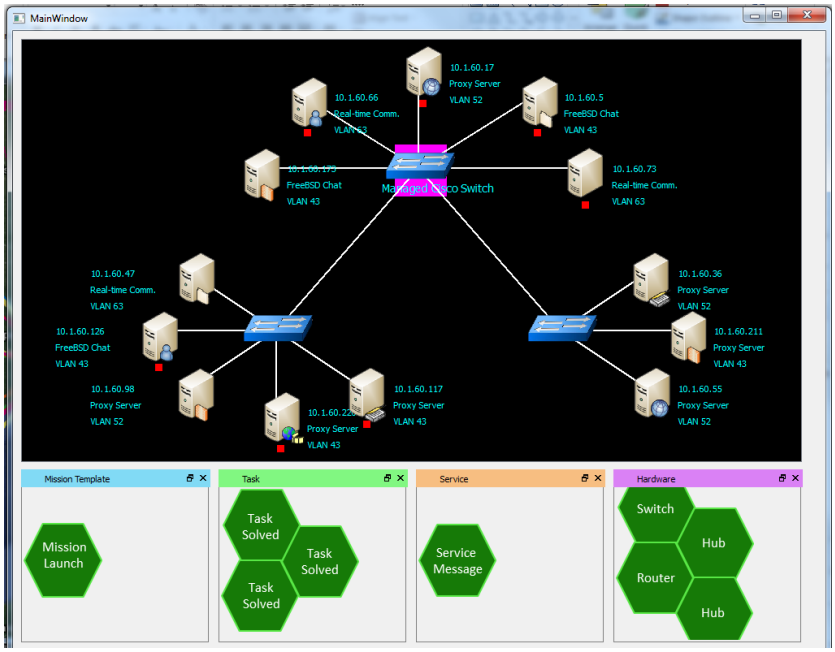


Figure 4-13 Cyber operation tasks: layer 3

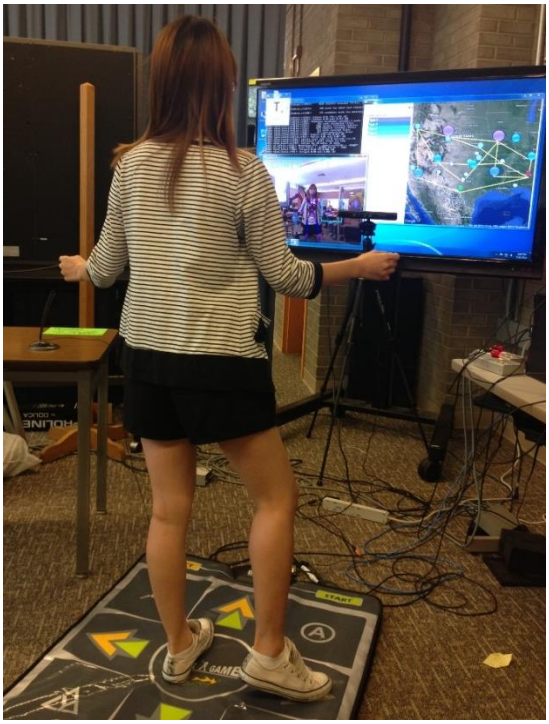


Figure 4-14 A prototyped multimodal interface used in a cyber-physical system

4.2.1 Design of Experiments

To accomplish this task, a total of 15 subjects were recruited, including 8 males and 7 females, all within the range of 20 – 30 years old. To validate the effectiveness of the multimodal system, the same task was completed, yet designed for a keyboard input system. Each subject interacted with both interfaces; subjects were allowed 10 trials for each interface. The summary of collected trials for each interface is presented in Table 4-2.

Table 4-2 Summary of collected trials for each interface

15 subjects (8 males, 7 females)	Embodied interaction based interface	Non-embodied interaction based interface
	10 trials / subject	10 trials / subject

During the experiment, observations were collected; these observations were further used to assess focus of attention through the use of the previously described BAN approach. The same performance metrics used in Case Study 1 were also utilized to assess user's performance in Case Study 2. Additionally, the subjects were also asked to complete a secondary task while solving the presented cyber operation task. The objective of the secondary task was to provide different method to assess attention, which will be further compared with the analytical approach presented in Chapter 3.4. Dual task is used as the baseline of attention here. The representative BAN obtained from Case Study 1 is used to assess the level of attention, and compare with the results of the secondary task.

The 1-back task [184] was used as the secondary task in Case Study 2. During the task, a T-like visual stimulus with two possible different shapes is presented to the participants (see Figure 4-15). It should be noted that the T sequence is randomly generated. The

participants have to distinguish whether the current T-stimulus is identical to the one presented at the last timestamp. The *hit rate*, represented by the percentage of accurate response of all shown responses, is used to measure the cognitive demand of the primary cyber-operation task. If the primary task requires high level of attention, the participant may frequently forget the previous stimulus, thus the hit rate will diminish with increasing attention-draw from the primary task.

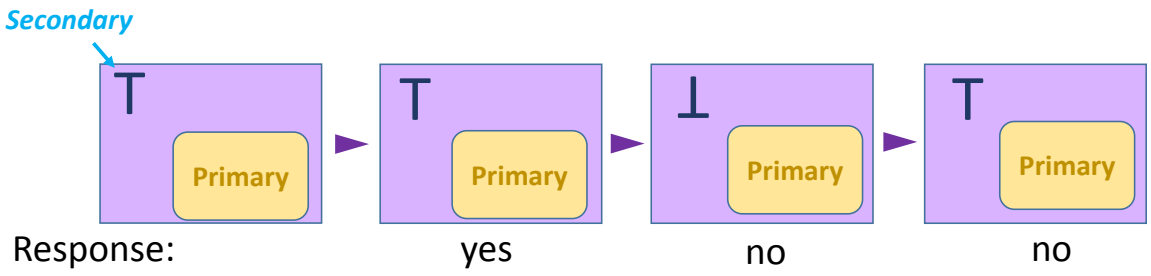


Figure 4-15 Example of stimulus sequence of a 1-back task and its correct responses for each T-like visual stimulus representation.

4.2.2 Results: BAN measure vs. Secondary Task measure

A total of 149 independent trials were collected from the multimodal interface. The representative BAN obtained from Case Study 1 was used to evaluate the focus of attention of those trials from Case Study 2. The secondary task (1-back task) that measured focus of attention was compared with the BAN approach. The results of these equivalence tests are summarized in Table 4-1. The null hypothesis of dissimilarity is not rejected at a difference value of 0.10 as the measured probability of attention from two approaches is different at a value of 0.10. However, when the difference between the measured levels of attention is 0.12, the null hypothesis is rejected. This rejection of the

null hypothesis means that the results of two approaches is equivalent enough (the difference is not significant). Figure 4-16 gives the bar graph displaying the mean and standard deviation of the two measurements.

Table 4-3 Statistical summary of equivalence tests for attention measure

Metric	Criterion	Dissimilarity	p-value
High level of attention (Probability)	0.10	Not rejected	0.113
High level of attention (Probability)	0.11	Not rejected	0.063
High level of attention (Probability)	0.12	Rejected	0.032*

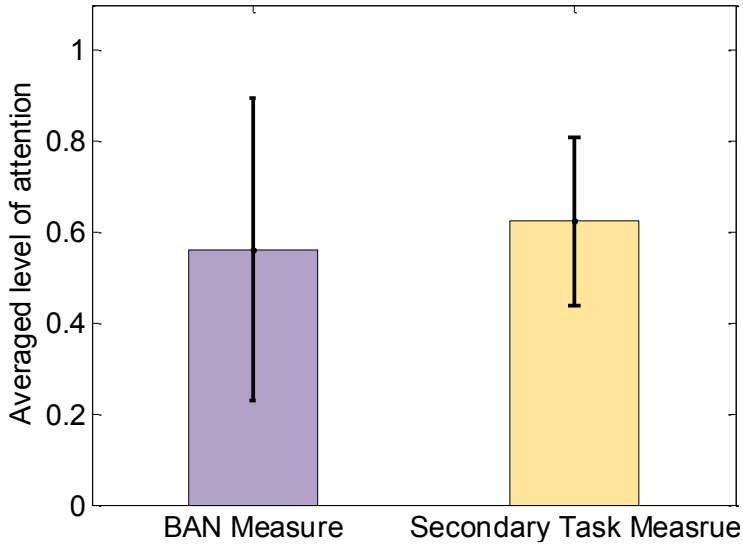


Figure 4-16 Level of assessed attention using two approaches

4.2.3 Results: Modalities Usage of the Multimodal Interface

Since the multimodal interface allows users to employ multiple interaction modalities during a task, the investigation of modality usage will be valuable for observing the relationships between a modality and the user's level of attention. Figure 4-17 shows the

3D scatter plot of level of attention (using BAN) vs. the fraction of times each modality used, given as a percentage. It can be noted from Figure 4-17 that the percentage of speech is nearly below 0.4 for to achieve higher focus of attention. By showing 2D plot with x-axis as the percentage of speech used and y-axis as the percentage of feet gestures used (see Figure 4-18), it is more evident that percentage of speech used is constrained in a range ([0.2, 0.4]). This would suggest that speech is not an optimal mode of interaction for spatial navigational tasks which require a greater level of attention.

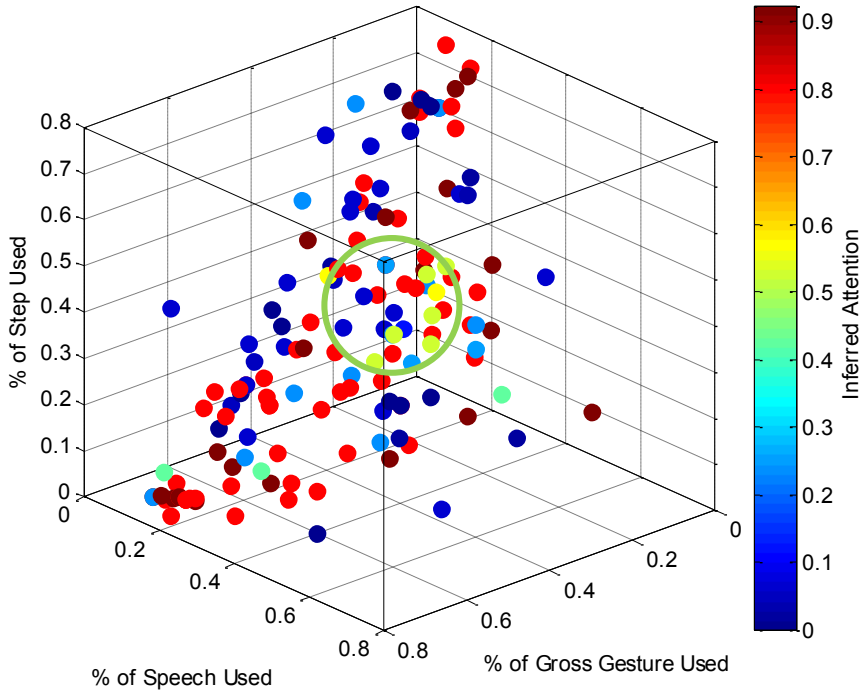


Figure 4-17 3D plot of inferred attention vs. the percentage of each modality used.

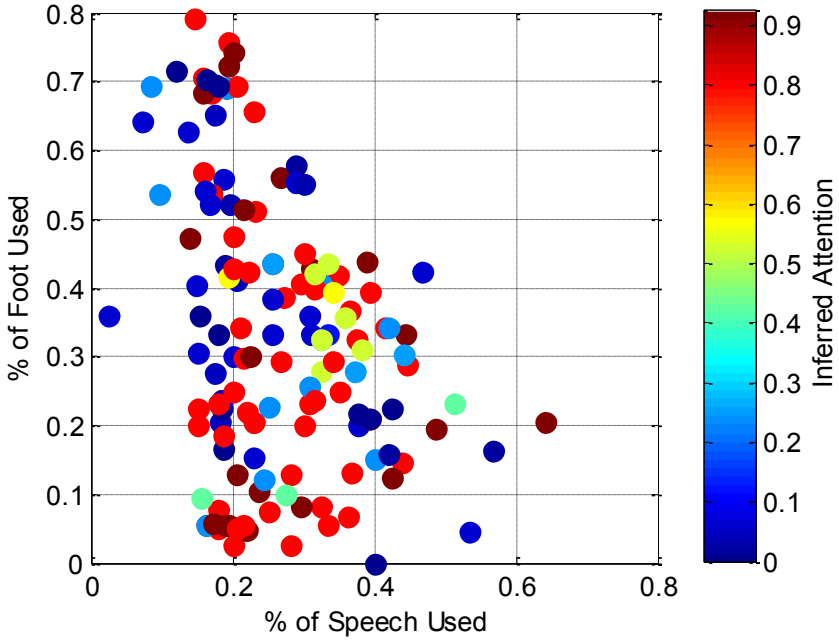


Figure 4-18 2D plot of inferred attention vs. the percentage of speech used

Another interesting finding is that employing all three modalities evenly is suggested to lead to a medium level of the focus of attention (the green cluster in the center of the graph). This can be explained as switching modalities may keep the user alert while somewhat preventing the user from fully focusing on the task. .

4.2.4 Results: Multimodal Interface vs. Keyboard Interface

The multimodal interface was compared with the keyboard interface to see whether embodied interaction results in better performance, compared to passive interaction. We collected a total of 149 independent trials for each interface. The metrics of task performance used in Case Study 1, including recognition rate of interaction, total task completion time, preparation time, and solution quality were measured during the experiment and further compared. In addition, a satisfaction survey was administered

after task completion. The relationship between the utility of interaction and performance metrics was determined through post-experiment data analysis, see Figure 4-19.

The results of a one-way ANOVA showed that keyboard interaction achieve higher recognition rate ($F(1,296)=145.803$, $p<0.0001$). Keyboard interaction also leads to spending less time on task ($F(1,296)=422.671$, $p < 0.0001$). However, the quality of solution of keyboard interaction is not significantly higher ($F(1,296)=1.804$, $p = 0.18$). Finally, the averaged level of a user's satisfaction of multimodal interaction is higher but not significantly ($F(1,296)=0.014$, $p = 0.907$).

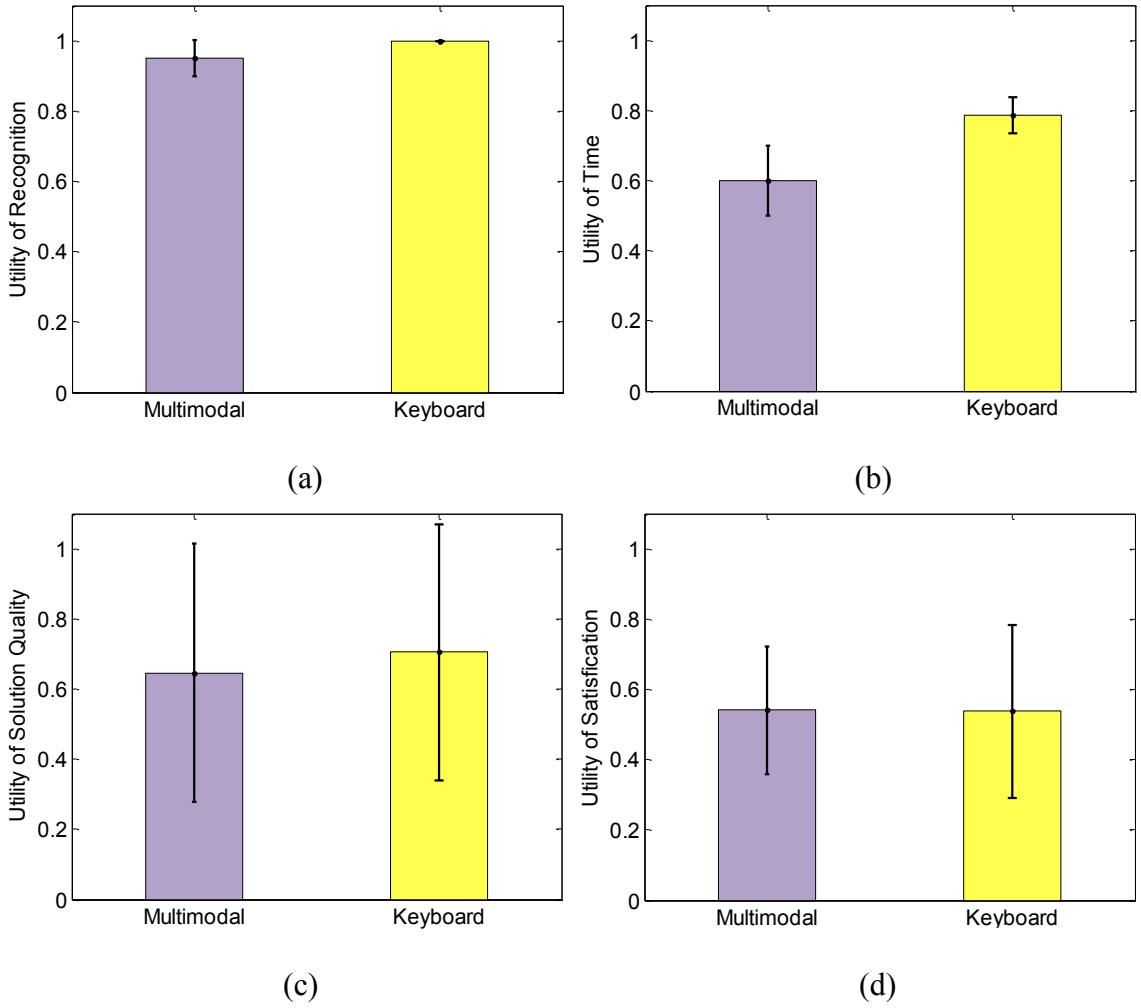


Figure 4-19 Expected utility vs. four performance metrics

However, an important finding is that, the error rate – errors occurred in task completion – when using embodied interaction in multimodal interfaces (4.37%) was significantly lower (p<0.05) than that of the traditional interaction (7.82%), shown in Figure 4-19. This is a clear advantage towards embodied interfaces.

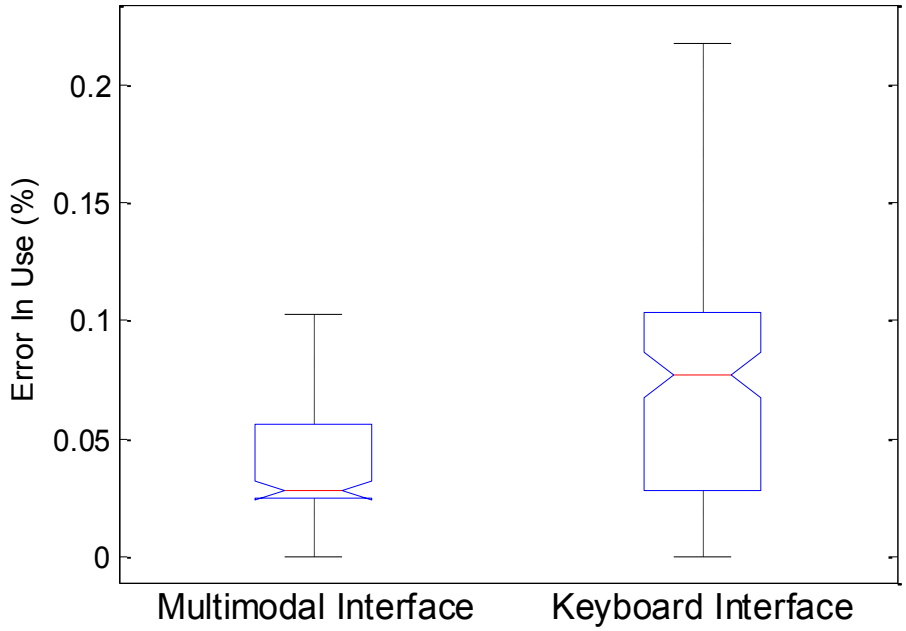


Figure 4-20 Comparison of error rate between two interfaces

The errors here refer to errors associated with failures to execute intent on the part of a user. For example, if the user apparently wishes to go to the node at the immediate right while pressing the key that will instead go to the bottom right, when there is no node connecting to the bottom right. Simply tapping on the keyboard requires the operator to remove his/her eyes from the region of interest to choose from a menu, thus increasing the errors in execution. Embodied interactions lead to fewer errors when compared to keyboard use, this may be due to fewer instances where the user breaks attention from the screen to observe the keyboard.

CHAPTER 5. DISCUSSION

This dissertation explores whether the performance and quality of the attempted solution to a decision making problem is affected by the way operators interact with visual data (e.g., physically, verbally, etc.). If this is true, it is important to determine the best combination of control and feedback modalities in terms of objective and subjective performance metrics, since they directly affect the solution. The scenario studied involves operators employing embodied interaction in solving spatially complex and time-sensitive problems. Through the use of cause-effect networks, derived from our proposed framework, five types of interaction modalities and two feedback modalities were cross-compared through a set of experiments in Case Study 1. This comparison was done between each combination of interaction and feedback modalities. The results showed that the use of feet on the dance pad controller led to better performance in all four metrics (recognition, time, quality of solution, and satisfaction) than using fine hand gestures (recognized through a data glove) for control and speech as feedback. Statistical analyses verified the existence of significant differences in the user utility ($p < .001$). It was observed that the dance pad led to excellent accuracy without affecting the operator's level of attention on the visualization surface. The foot gesture coupled with visual feedback modality achieved high accuracy, saved task completion time, and mitigated users' distraction. A possible explanation is that foot gestures are a suitable form of

interaction to convey intent in such spatial navigational task. The stepping movement performed during the interactions may promote mental simulation of walking, aiding navigational actions. For similar reasons, speech becomes an inappropriate selection for interaction as it does not enable this specific form of mental simulation. Another possible explanation for the inadequacy of speech-based interaction is that this particular mode requires retention and recollection of the specific vocabulary permitting interaction.

There is a key question that emerges from this research that is how we can tell that the attentional level computed are the real ones exhibited by the user, without using intrusive methods. For this, the proposed approach of measuring level of attention has been compared to the secondary task approach via embodied interaction based multimodal interface. This was one of the objectives of Case Study 2, in which it was showed that the proposed BAN is able to measure level of attention consistently with state-of-the-art approach (secondary task measure) in an objective, quantitative, and non-intrusive manner.

5.1 Discussion: Bayesian Attentional Network

The BAN model proposed also yield some more provocative findings. Careful analysis of the adjacency matrix (obtained through the CMM method proposed (Figure 4-6)), shows that there was a certain agreement around the cause-effect between attention and torso orientation (expressed through the edge connecting the nodes “attention” and “torso orientation”) in the candidate BANs. We believe that the reason for this is that the use of foot-based interaction allows the operator’s torso to face towards the screen while simultaneously permitting the operator’s eyes to continuously focus on the display. This,

in turn, leads to an increased use of his/her body during the tasks regardless of finger configurations or speech commands.

An important question is whether the evidence related to physical actions added information to the BAN beyond the information conveyed through the variables speed and accuracy, which are widely used in cognitive psychology (we referred to those variables with the names of “inter-command elapsed time” and “error in use”, respectively) to assess attention. In this respect, we found that the probability for high level of attention was not consistent with our findings when using only speed and accuracy. This is an indication that indeed the physical expressions added another “layer” of evidence for assessing attention, not previously explored in the area of cognitive psychology.

While the utilization of BANs for modeling the level of attention shows high face validity, it does possess some limitations. Firstly, the BANs cannot capture instantaneous attention for a given time, and instead it is possible only to see the cumulative effect of the observed quantity at the time that a command is evoked. This is why probabilities at the evidence nodes are obtained only after the operator evokes a command. A possible solution for this is to adopt a Dynamic Bayesian Network to model the continuum of attention span between consecutive commands. The second problem observed is related to the complexity associated with running the evolutionary approach. Although the scores of evolutionary approach improved significantly (25.08% at most, and 9.77% at least) from their initial values, a single implementation of this algorithm can take around 10 hours using a computer equipped with 8-core processor. Therefore the number of initial

solutions was limited to only five. Finding more effective ways to compute the scoring metric would speed this process, eventually leading to more attractive solutions.

It was found that the measure of attention using the BAN approach and the secondary task approach lead to consistency with a criteria in Case Study 2. The criteria was empirically selected by the author such that a difference of 0.12 in the level of probability between two measures is considered insignificant. To test whether this criteria is good, a power analysis was conducted with an effect size of 0.12. This analysis showed that the power is 0.959 with 149 number of trials and an effect size of 0.12. The power is high thus it can be claimed that the BAN measure is consistent with state-of-the-art measures when the probability difference is around 0.12.

However, the secondary task measure still required the users to react to visual stimulus and determine their response. It is still possible that user made incorrect decisions while fully focusing on the secondary task. In this case, the secondary task still cannot be used as an evaluation of the absolute ground truth, or the baseline for comparisons. The dual task method is designed to determine level of attention directed to a primary task by measuring the level of interference generated by a secondary task. We have developed an alternative method for the determination of attention which does not require the generation of interference between two tasks. Within Case Study 2, we are concerned with a primary, real-time task, the manipulation of cyber-physical systems. The use of dual-task measures would cause interference with the main task, observed by lower performance metrics. Our BAN method more directly measures level of attention; this fact was validated by application of and comparison to the dual-task method. Thus, our BAN method is applicable to the single-task or the dual-task scenario. Other

physiological measures (such as eye tracking, capable of measuring lengthy distance between eyes and the screen, or electroencephalography (EEG)) can be adopted to measure the behavior response of a user. Table 5-1 summarizes the comparison of previous work and our research about assessing and reasoning user's state of attention.

Table 5-1 Summary of attention-supported user interface and comparisons with our work

System	Key property	Method/Device	Reference
GAZE	Used eye trackers to assess visual awareness among users in a group	Virtual Reality Modeling/Eye tracker	Vertegaal, 1999
iTourist	Exploited the user's gaze pattern to help city trip planning	Eye tracker	Qvarfordt and Zhai, 2005
Priorities System	Predict the urgency of incoming emails and decide the most appropriate time of notification.	Support Vector Machine	Horvitz et al., 1999
COORDINATE	Used Bayesian learning and inference to predicts the user's presence and availability	Bayesian network	Horvitz et al., 2002
Notification Platform	Probabilistically observed a user's level of attention based on user's perceptual evidence (gaze, utterance) scheduled activities	Bayesian network, Hidden Markov Models	Horvitz et al., 2003
Our work: Bayesian Attentional Network (BAN)	Assessed user's level of attention while he adopted embodied interactions (gestures, utterance, and body stance) to navigate and interact with visual information	Bayesian network, Utility theory	Li and Wachs, 2014

5.2 Discussion: Task Performance

The performance metrics used in Case Study 2 give evidence suggesting that keyboard interface achieves higher accuracy ($p < .001$) and less time spent on a task ($p < .001$). All of the subjects had, at minimum, 10 years of experience on keyboard interfaces. Even while undertaking a novel task the subjects were extremely familiar with the keyboard interface. On the other hand, most of the subjects had no experience in using embodied interaction to facilitate interaction with the computing devices. Portions of time spent on the multimodal system were likely used for exploring and familiarizing the user with the novel interface. This familiarization time could explain the lengthier time associated with task completion for unfamiliar modalities; this time being minimal for the familiar keyboard modality.

It was also noted that keyboard usage has higher execution errors than does the multimodal interface (4.37% vs. 7.82%, $p<.05$) even though users are more familiar with the keyboard. This finding can be explained and is consistent with the claim that traditional interfaces are limited in dealings with complex applications or spatial data [1], [2], and are not suitable interfaces for diverse display and analysis environments. Traditional interfaces also create a gap between a user's intent and its execution, and execution errors increase due to the existence of this gap. Foglia and Wilson [185] argued that spatial concepts (such as 'front', 'back', 'up', and 'down') both arise from and are articulated by our particular body shape as well as the manner in which we navigate our bodies within space. Table 5-2 compares systems shown in the literature that involve embodied interaction in the completion of spatial tasks.

Table 5-2 Summary of previous embodied interaction based interface and comparisons with our work

	Interacting scenario	Method/Device	Modalities
Pakkanen and Raisamo, 2004	Allowed users to manipulate a graphical user interface by the foot in different non-accurate spatial tasks	- Trackball	- Foot gesture - Hand gesture
Schöning et al., 2009	Applied multi-touch hand gestures and foot gestures to interact with a Geographic Information System (GIS) on a large-scale interactive screen	- Multi-touch surface - Wii Balance Board	- Foot gesture - Hand gesture
Daiber et al., 2011	Presented a multi-modal interaction with a GIS on large-scale displays by using multi-hand touch, foot and gaze input	- Multi-touch surface - Wii Balance Board - Eye tracker	- Foot gesture - Hand gesture - Gaze
Göbel et al., 2013	Presented a gaze-supported foot interaction to support exploration, selection, and modification task in a GIS.	- Fanatec CSR Elite1 foot pedal - Custom-made foot-joystick and foot-rocker - Eye tracker	- Foot gesture - Gaze
Li and Wachs 2014 (Our work)	Developed a multi-modal embodied interaction system to navigate spatial decision making problems (TSP, Cyber-physical system)	- Kinect - Dance pad - Wii Balance Board - 5DT data glove - Microphone	- Foot gesture - Hand gesture - Speech - Body stance

However, embodied interfaces do not significantly outperform the keyboard interfaces in terms of the metrics we adopted. The reasons for this can be the lack of complexity in

data dimensions in our studies. The cyber-physical operations are shown in a 2D plain graph in our implementations during Case Study 2, and thus the advantage of embodied interaction may not be fully demonstrated. For example, performing a hand gesture forward or backward allows the users to navigate the depth dimension more intuitively than does the use of keyboard. Future work can be extended by constructing the 3D map where the locations of bases are on the surface of a manifold (such as Google Earth), and allowing 3D navigation instead of 2D.

In terms of the utility function, we assumed that there exists a direct relationship between performance metrics and any benefits/costs associated with interaction. While this assumption simplifies our problem, other functions could be plugged into our framework easily without modifying any of the principles underlying its design and theory. Also, relative importance was assumed to be equal between the various metrics. An alternative approach would be to use the Analytic Hierarchy Process (AHP) to determine more realistic weights corresponding to the preferences and priorities of an operator.

CHAPTER 6. CONCLUSIONS AND FUTURE WORK

In this dissertation, we proposed to study the effect of embodied interaction during the solution of complex and time-sensitive decision making problems. A method linking an operator's interaction utility, inference and reasoning for the assessment of the level of an operator's attention was presented herein. The approach discussed consisted of developing a new methodology to infer user's attention based on disparate raw signals from multiple channels, and calculating the utility of embodied interaction effectively through Bayesian networks. We call these networks Bayesian Attentional Networks (BANs). BANs are structures describing the cause-effect relationship between operators' level of attention, physical action and decision-making in spatial temporally complex and time-sensitive scenarios. A number of metrics were developed for expressing the benefits and costs of different control and feedback modalities. An enhanced combination of control and feedback was determined using objective and subjective metrics. The proposed framework considers both the operator's knowledge and a biologically inspired method to compute the BAN (associated with the highest objective function). This BAN was obtained through the innovative CMM method. This method automatically creates a representative BAN based on the consensus level among the proposed candidate solutions. This approach is an extrapolation of the well-known RANSAC method used within statistics. RANSAC's basic concept consists of selecting the subset of instances

(in our case BANs) that can best explain the model and its parameters. The candidate models found were those that met the maximum agreement among the inliers (note, consensus is a special case of overall agreement). The resulting network obtained through the CMM method, explains why level of attention not only affects the physical action but also the task performance. Leveraging on this approach to assess attention levels, utility theory was then used to express the trade-off between benefits and cost associated with various performances metrics. To summarize, we presented three main contributions: (1) the BAN that builds a cause-effect relationship between physical actions, level of attention and decision-making; (2) the CMM which consolidates BANs obtained from different sources; and (3) the utility function that determines the most suitable combination of interaction modalities and feedback so as to enhance operator's task performance.

The our Case Study 1, results showed that the dance pad controller allows operators to explore an image (as if they were "walking through") while keeping their eyes focused on the screen, thus increasing attention and thus task performance. The embodied interaction based multimodal interface was also integrated within a cyber-physical threat resolution system in Case Study 2, with the objective of decreasing task completion errors. Measurements taken during this task were also used to validate that the BANs can appropriately infer an operator's attention while using an embodied interaction based multimodal interface. Table 6-1 summarizes the objective and insights of Case Studies in this dissertation, and the relations to our research questions.

Table 6-1 Summary of two Case Studies and their relation to the research questions

	Research Question	Case Study
RQ1	What is the optimal combination of interaction modalities and feedback that lead to the best task performance (among the alternatives studied)?	Case Study 1: -Discover the optimal combination interaction and feedback modalities by comparing various metric. - The combination of foot gestures with visual feedback resulted in the best task performance, including accuracy, shorter time, better quality of solution, and user's experience.
RQ2	Which benefits are offered by embodied interaction over those offered by non-embodied interaction method during the completion of spatial navigational scenarios?	Case Study 2: - Compare the embodied based interaction with non-embodied based interaction using various metrics. - Embodied interaction based interaction is outperforming the non-embodied interaction with the benefit of reducing execution errors.

Future work will involve testing this approach with a larger dimensional decision making problem. For example, integrating applications used to interact with 3D visualization of cyber-physical operations, and extending the evidence nodes to include additional sensed information, such as force-feedback and gaze direction. In such scenarios an important question to be addressed is how generalizable the derived BAN is across several spatial navigational tasks. That is, while attention was inferred though the CMM model (proposed in the task of study - the TSP), it is not clear how well it can infer operators'

attention in other tasks requiring decision-making. Wilson and Golonka [186] sustain that embodied cognition provides the solutions that solve specific tasks, but not general problems. Thus, key challenges involve determining whether these embodied cognition based solutions can be applied to common tasks, such as non-spatial navigation ones. Further effort is needed to address this problem.

REFERENCES

REFERENCES

[1] B. Lee, P. Isenberg, N. H. Riche, and S. Carpendale, "Beyond Mouse and Keyboard: Expanding Design Considerations for Information Visualization Interactions," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2689–2698, Dec. 2012.

[2] J. Schöning, F. Daiber, A. Krüger, and M. Rohs, "Using Hands and Feet to Navigate and Manipulate Spatial Data," in *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2009, pp. 4663–4668.

[3] M. C. Wright, J. M. Taekman, and M. R. Endsley, "Objective measures of situation awareness in a simulated medical environment," *Qual. Saf. Health Care*, vol. 13, no. suppl 1, pp. i65–i71, Oct. 2004.

[4] L. Weschke and G. Börman, "Pushing Towards Embodied Interactions," Sep. 2010.

[5] G. Lakoff and M. Johnson, "The Metaphorical Structure of the Human Conceptual System," *Cogn. Sci.*, vol. 4, no. 2, pp. 195–208, 1980.

[6] S. R. Klemmer, B. Hartmann, and L. Takayama, "How Bodies Matter: Five Themes for Interaction Design," in *Proceedings of the 6th Conference on Designing Interactive Systems*, New York, NY, USA, 2006, pp. 140–149.

[7]H. Bekkering and S. F. W. Neggers, "Visual search is modulated by action intentions," *Psychol. Sci.*, vol. 13, no. 4, pp. 370–374, Jul. 2002.

[8] E. Balcetis and D. Dunning, "Cognitive dissonance and the perception of natural environments," *Psychol. Sci.*, vol. 18, no. 10, pp. 917–921, Oct. 2007.

[9] A. M. Glenberg, D. Havas, R. Becker, and M. Rinck, "Grounding Language in Bodily States: The Case for Emotion," in *Grounding Cognition*, Cambridge University Press, 2005.

[10] C. L. Scott, R. J. Harris, and A. R. Rothe, "Embodied Cognition Through Improvisation Improves Memory for a Dramatic Monologue," *Discourse Process.*, vol. 31, no. 3, pp. 293–305, 2001.

[11] R. Nemirovsky, C. Tierney, and T. Wright, "Body Motion and Graphing," *Cogn. Instr.*, vol. 16, no. 2, pp. 119–172, 1998.

[12] S. Pirie and T. Kieren, "Growth in Mathematical Understanding: How Can We Characterise It and How Can We Represent It?," *Educ. Stud. Math.*, vol. 26, pp. 165–90, Jan. 1994.

[13] S. Mann, R. E. Janzen, R. Lo, and J. Fung, "Non-electrophonic Cyborg Instruments: Playing on Everyday Things As if the Whole World Were One Giant Musical Instrument," in *Proceedings of the 15th International Conference on Multimedia*, New York, NY, USA, 2007, pp. 932–941.

[14] S. Fels, “Intimacy and Embodiment: Implications for Art and Technology,” in *Proceedings of the 2000 ACM Workshops on Multimedia*, New York, NY, USA, 2000, pp. 13–16.
[15] P. Dourish, *Where the Action is: The Foundations of Embodied Interaction*. MIT Press, 2004.
[16] A. Clark, *Being There: Putting Brain, Body, and World Together Again*. MIT Press, 1997.
[17] L. W. Barsalou, “Grounded Cognition,” *Annu. Rev. Psychol.*, vol. 59, no. 1, pp. 617–645, 2008.
[18] J. B. Black, “An Embodied/Grounded Cognition Perspective on Educational Technology,” in *New Science of Learning*, M. S. Khine and I. M. Saleh, Eds. Springer New York, 2010, pp. 45–52.
[19] A. Segal, “Do Gestural Interfaces Promote Thinking? Embodied Interaction: Congruent Gestures and Direct Touch Promote Performance in Math,” COLUMBIA UNIVERSITY, 2011.
[20] K. Johnson, J. Pavleas, and J. Chang, “Kinecting to Mathematics through Embodied Interactions,” *Computer*, vol. 46, no. 10, pp. 101–104, Oct. 2013.
[21] M. Chu and S. Kita, “The nature of gestures’ beneficial role in spatial problem solving,” *J. Exp. Psychol. Gen.*, vol. 140, no. 1, pp. 102–116, 2011.
[22] M. Wilson, “Six views of embodied cognition,” *Psychon. Bull. Rev.*, vol. 9, no. 4, pp. 625–636, Dec. 2002.
[23] A. B. Hostetter and M. W. Alibali, “Visible embodiment: Gestures as simulated action,” *Psychon. Bull. Rev.*, vol. 15, no. 3, pp. 495–514, Jun. 2008.
[24] L. Aziz-Zadeh, S. M. Wilson, G. Rizzolatti, and M. Iacoboni, “Congruent embodied representations for visually presented actions and linguistic phrases describing actions,” *Curr. Biol. CB*, vol. 16, no. 18, pp. 1818–1823, Sep. 2006.
[25] R. Ball, C. North, and D. A. Bowman, “Move to Improve: Promoting Physical Navigation to Increase User Performance with Large Displays,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2007, pp. 191–200.
[26] R. Ball and C. North, “Realizing embodied interaction for visual analytics through large displays,” *Comput. Graph.*, vol. 31, no. 3, pp. 380–400, Jun. 2007.
[27] C. Hummels, K. C. Overbeeke, and S. Klooster, “Move to Get Moved: A Search for Methods, Tools and Knowledge to Design for Expressive and Rich Movement-based Interaction,” *Pers. Ubiquitous Comput*, vol. 11, no. 8, pp. 677–690, Dec. 2007.
[28] D. Wigdor and D. Wixon, *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*, 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.
[29] M. G. Jacob, Y.-T. Li, G. A. Akingba, and J. P. Wachs, “Collaboration with a robotic scrub nurse,” *Commun ACM*, vol. 56, no. 5, pp. 68–75, May 2013.
[30] A. van Dam, “Post-WIMP User Interfaces,” *Commun ACM*, vol. 40, no. 2, pp. 63–67, Feb. 1997.
[31] C. D. Wickens, *Applied Attention Theory*. Boca Raton: CRC Press, 2007.
[32] W. James, *The Principles of Psychology*. H. Holt, 1918.

[33] E. Horvitz, A. Jacobs, and D. Hovel, “Attention-sensitive alerting,” in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, San Francisco, CA, USA, 1999, pp. 305–313.
[34] D. Kahneman, *Attention and Effort*. Prentice-Hall Inc., 1973.
[35] B. Bailey, J. Konstan, and J. Carlis, “The effects of interruptions on task performance, annoyance, and anxiety in the user interface,” presented at the Proceedings of INTERACT, 2001, vol. 1, pp. 593–601.
[36] M. A. Recarte and L. M. Nunes, “Mental workload while driving: effects on visual search, discrimination, and decision making,” *J. Exp. Psychol. Appl.*, vol. 9, no. 2, pp. 119–137, Jun. 2003.
[37] H. Pashler, “Dual-task interference in simple tasks: Data and theory,” *Psychol. Bull.*, vol. 116, no. 2, pp. 220–244, 1994.
[38] D. L. Damos, *Multiple Task Performance*. CRC Press, 1991.
[39] L. Gugerty and M. Falzetta, “Using an Event-Detection Measure to Assess Drivers’ Attention and Situation Awareness,” *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 49, no. 22, pp. 2004–2008, Sep. 2005.
[40] A. Poole and L. J. Ball, “Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future,” in *Prospects”, Chapter in C. Ghaoui (Ed.): Encyclopedia of Human-Computer Interaction. Pennsylvania: Idea Group, Inc*, 2005.
[41] K. Moore and L. Gugerty, “Development of a Novel Measure of Situation Awareness: The Case for Eye Movement Analysis,” *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 54, no. 19, pp. 1650–1654, Sep. 2010.
[42] J. E. Richards and B. J. Casey, “Heart Rate Variability During Attention Phases in Young Infants,” *Psychophysiology*, vol. 28, no. 1, pp. 43–53, Jan. 1991.
[43] W. A. Johnston and V. J. Dark, “Selective Attention,” *Annu. Rev. Psychol.*, vol. 37, no. 1, pp. 43–75, 1986.
[44] E. C. Cherry, “Some Experiments on the Recognition of Speech, with One and with Two Ears,” *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.
[45] A. M. Bonnel and E. R. Hafter, “Divided attention between simultaneous auditory and visual signals,” *Percept. Psychophys.*, vol. 60, no. 2, pp. 179–190, Feb. 1998.
[46] S. T. Iqbal, Y.-C. Ju, and E. Horvitz, “Cars, calls, and cognition: investigating driving and divided attention,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2010, pp. 1281–1290.
[47] K. A. Brookhuis, G. de Vries, and D. de Waard, “The effects of mobile telephoning on driving performance,” *Accid. Anal. Prev.*, vol. 23, no. 4, pp. 309–316, Aug. 1991.
[48] A. J. McKnight and A. S. McKnight, “The effect of cellular phone use upon driver attention,” *Accid. Anal. Prev.*, vol. 25, no. 3, pp. 259–265, Jun. 1993.
[49] V. Briem and L. R. Hedman, “Behavioural effects of mobile telephone use during simulated driving,” *Ergonomics*, vol. 38, no. 12, pp. 2536–2562, 1995.
[50] M. Corbetta and G. L. Shulman, “Control of goal-directed and stimulus-driven attention in the brain,” *Nat. Rev. Neurosci.*, vol. 3, no. 3, pp. 201–215, Mar. 2002.

[51] T. J. Buschman and E. K. Miller, “Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices,” *Science*, vol. 315, no. 5820, pp. 1860–1862, Mar. 2007.
[52] M. M. Chun and J. Wolfe, *Visual attention*. Oxford, UK: Blackwell, 2001.
[53] Y. Fang, W. Lin, C. T. Lau, and B.-S. Lee, “A visual attention model combining top-down and bottom-up mechanisms for salient object detection,” in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 1293 –1296.
[54] H. L. O’Brien and E. G. Toms, “What is user engagement? A conceptual framework for defining user engagement with technology,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, no. 6, pp. 938–955, 2008.
[55] C. Peters, G. Castellano, and S. de Freitas, “An exploration of user engagement in HCI,” in *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*, New York, NY, USA, 2009, pp. 9:1–9:3.
[56] R. Vertegaal, “Attentive User Interfaces,” *Commun. ACM*, vol. 46, no. 3, pp. 30–33, 2003.
[57] C. Roda and J. Thomas, “Attention aware systems: Theories, applications, and research agenda,” *Comput. Hum. Behav.*, vol. 22, no. 4, pp. 557–587, Jul. 2006.
[58] C. Speier, I. Vessey, and J. S. Valacich, “The Effects of Interruptions, Task Complexity, and Information Presentation on Computer-Supported Decision-Making Performance,” *Decis. Sci.*, vol. 34, no. 4, pp. 771–797, 2003.
[59] D. C. McFarlane and K. A. Latorella, “The Scope and Importance of Human Interruption in Human-computer Interaction Design,” *Hum-Comput Interact*, vol. 17, no. 1, pp. 1–61, Mar. 2002.
[60] A. Oulasvirta, S. Tamminen, V. Roto, and J. Kuorelahti, “Interaction in 4-second Bursts: The Fragmented Nature of Attentional Resources in Mobile HCI,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2005, pp. 919–928.
[61] D. Goldman, “Google unveils ‘Project Glass’ virtual-reality glasses,” *CNNMoney*. [Online]. Available: http://money.cnn.com/2012/04/04/technology/google-project-glass/index.htm. [Accessed: 20-Feb-2014].
[62] S. Zulkernain, P. Madiraju, S. I. Ahamed, and K. Stamm, “A Mobile Intelligent Interruption Management System,” *J. Univers. Comupter Sci.*, Jan. 2010.
[63] R. Vertegaal and J. S. Shell, “Attentive user interfaces: the surveillance and sousveillance of gaze-aware objects,” *Soc. Sci. Inf.*, vol. 47, no. 3, pp. 275–298, Sep. 2008.
[64] R. A. Bolt, “Conversing with computers,” *Technol. Rev.*, vol. 6, no. 2, Mar. 1985.
[65] R. J. K. Jacob, “Virtual environments and advanced interface design,” W. Barfield and T. A. Furness,III, Eds. New York, NY, USA: Oxford University Press, Inc., 1995, pp. 258–288.
[66] S. Zhai, C. Morimoto, and S. Ihde, “Manual and gaze input cascaded (MAGIC) pointing,” in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, New York, NY, USA, 1999, pp. 246–253.

[67] I. Starker and R. A. Bolt, “A gaze-responsive self-disclosing display,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 1990, pp. 3–10.

[68] P. Qvarfordt and S. Zhai, “Conversing with the user based on eye-gaze patterns,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2005, pp. 221–230.

[69] R. Vertegaal, “The GAZE groupware system: mediating joint attention in multiparty communication and collaboration,” in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, New York, NY, USA, 1999, pp. 294–301.

[70] R. Vertegaal, I. Weevers, C. Sohn, and C. Cheung, “GAZE-2: conveying eye contact in group video conferencing using eye-controlled camera direction,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2003, pp. 521–528.

[71] “The Samsung Galaxy S4 and eye-tracking,” *Electronics News*. [Online]. Available: http://www.electronicsnews.com.au/Features/The-Samsung-Galaxy-S4-and-eye-tracking. [Accessed: 25-Apr-2013].

[72] E. Horvitz, P. Koch, C. M. Kadie, and A. Jacobs, “Coordinate: probabilistic forecasting of presence and availability,” in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, San Francisco, CA, USA, 2002, pp. 224–233.

[73] E. Horvitz, C. Kadie, T. Paek, and D. Hovel, “Models of Attention in Computing and Communication: From Principles to Applications,” *Commun ACM*, vol. 46, no. 3, pp. 52–59, Mar. 2003.

[74] E. Horvitz and J. Apacible, “Learning and reasoning about interruption,” in *Proceedings of the 5th international conference on Multimodal interfaces*, New York, NY, USA, 2003, pp. 20–27.

[75] “ACM SIGCHI curricula for human-computer interaction,” ACM, New York, NY, USA, 1992.

[76] J. Grudin, “Three faces of human-computer interaction,” *IEEE Ann. Hist. Comput.*, vol. 27, no. 4, pp. 46 – 62, Dec. 2005.

[77] J. M. Carroll, “HUMAN-COMPUTER INTERACTION: Psychology as a Science of Design,” *Annu. Rev. Psychol.*, vol. 48, no. 1, pp. 61–83, 1997.

[78] G. Lindgaard and C. Dudek, “What is this evasive beast we call user satisfaction?,” *Interact. Comput.*, vol. 15, no. 3, pp. 429–452, Jun. 2003.

[79] F. Karray, M. Alemzadeh, J. A. Saleh, and M. N. Arab, “Human-Computer Interaction: Overview on State of the Art,” *Int. J. Smart Sens. Intell. Syst.*, vol. 1, no. 1, pp. 137–159, Mar. 2008.

[80] R. Ballagas, J. Borchers, M. Rohs, and J. G. Sheridan, “The smart phone: a ubiquitous input device,” *IEEE Pervasive Comput.*, vol. 5, no. 1, pp. 70 – 77, Mar. 2006.

[81] Y. Nakhimovsky, D. Eckles, and J. Riegelsberger, “Mobile user experience research: challenges, methods & tools,” in *CHI ’09 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2009, pp. 4795–4798.

[82] K. Montague, V. L. Hanson, and A. Cobley, “Designing for individuals: usable touch-screen interaction through shared user models,” in *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, New York, NY, USA, 2012, pp. 151–158.

[83] S. Kim, S.-H. Kim, and H.-G. Cho, “Developing a system for searching a shop name on a mobile device using voice recognition and GPS information,” in *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*, New York, NY, USA, 2012, pp. 27:1–27:8.

[84] “Samsung’s new TV remote has touch, motion, voice control,” *CNET*. [Online]. Available: http://www.cnet.com/news/samsungs-new-tv-remote-has-touch-motion-voice-control/.

[85] T. Geller, “Talking to Machines,” *Commun ACM*, vol. 55, no. 4, pp. 14–16, Apr. 2012.

[86] I. Poupyrev and S. Maruyama, “Tactile interfaces for small touch screens,” in *Proceedings of the 16th annual ACM symposium on User interface software and technology*, New York, NY, USA, 2003, pp. 217–220.

[87] L. E. Sibert and R. J. K. Jacob, “Evaluation of Eye Gaze Interaction,” 2000, pp. 281–288.

[88] S. Brewster and L. M. Brown, “Tactons: structured tactile messages for non-visual information display,” in *Proceedings of the fifth conference on Australasian user interface - Volume 28*, Darlinghurst, Australia, Australia, 2004, pp. 15–23.

[89] A. Jaimes and N. Sebe, “Multimodal human-computer interaction: A survey,” *Comput Vis Image Underst*, vol. 108, no. 1–2, pp. 116–134, Oct. 2007.

[90] H. Bullinger, J. Ziegler, W. Bauer, and F. Iao, “Intuitive human-computer interaction—Toward a user-friendly information society,” *Int J Hum.-Comput Interact.*, pp. 1–23, 2002.

[91] D. Gouin and V. Lavigne, “Trends in Human-Computer Interaction to Support Future Intelligence Analysis Capabilities,” Jun. 2011.

[92] N. Beringer, U. Kartal, K. Louka, F. Schiel, and U. Türk, “PROMISE - A Procedure for Multimodal Interactive System Evaluation,” *Multimodal Resour. Multimodal Syst. Eval. Workshop Program*, 2002.

[93] I. Wechsung, K.-P. Engelbrecht, S. Schaffer, J. Seebode, F. Metze, and S. Möller, “Usability Evaluation of Multimodal Interfaces: Is the Whole the Sum of Its Parts?,” in *Human-Computer Interaction. Novel Interaction Methods and Techniques*, J. A. Jacko, Ed. Springer Berlin Heidelberg, 2009, pp. 113–119.

[94] D. England, J. Fantauzzacoffin, N. Bryan-Kinns, C. Latulipe, L. Candy, and J. Sheridan, “Digital art: evaluation, appreciation, critique (invited SIG),” in *Proceedings of the 2012 ACM annual conference extended abstracts on Human Factors in Computing Systems Extended Abstracts*, New York, NY, USA, 2012, pp. 1213–1216.

[95] A. Pentland, “Perceptual user interfaces: perceptual intelligence,” *Commun ACM*, vol. 43, no. 3, pp. 35–44, Mar. 2000.

[96] A. Legin, A. Rudnitskaya, B. Seleznev, and Y. Vlasov, “Electronic tongue for quality assessment of ethanol, vodka and eau-de-vie,” *Anal. Chim. Acta*, vol. 534, no. 1, pp. 129–135, Apr. 2005.

[97] R. Sharma, V. I. Pavlovic, and T. S. Huang, “Toward multimodal human-computer interface,” *Proc. IEEE*, vol. 86, no. 5, pp. 853 –869, May 1998.
[98] S. Goldin-Meadow, “The role of gesture in communication and thinking,” *Trends Cogn. Sci.*, vol. 3, no. 11, pp. 419–429, Nov. 1999.
[99] M. A. Moni and A. B. M. S. Ali, “HMM based hand gesture recognition: A review on techniques and approaches,” in *2nd IEEE International Conference on Computer Science and Information Technology, 2009. ICCSIT 2009*, 2009, pp. 433 – 437.
[100] V. I. Pavlovic, R. Sharma, and T. S. Huang, “Visual interpretation of hand gestures for human-computer interaction: a review,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 677 –695, Jul. 1997.
[101] M. R. Morris, A. Huang, A. Paepcke, and T. Winograd, “Cooperative gestures: multi-user gestural interactions for co-located groupware,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2006, pp. 1201–1210.
[102] R. Poppe, R. Rienks, and B. van Dijk, “Evaluating the Future of HCI: Challenges for the Evaluation of Emerging Applications,” in *Artifical Intelligence for Human Computing*, T. S. Huang, A. Nijholt, M. Pantic, and A. Pentland, Eds. Springer Berlin Heidelberg, 2007, pp. 234–250.
[103] S. S. Rautaray and A. Agrawal, “Interaction with virtual game through hand gesture recognition,” in *2011 International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT)*, 2011, pp. 244 –247.
[104] S. Mitra and T. Acharya, “Gesture Recognition: A Survey,” *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 37, no. 3, pp. 311 –324, May 2007.
[105] J. P. Wachs, M. Kölsch, H. Stern, and Y. Edan, “Vision-based hand-gesture applications,” *Commun ACM*, vol. 54, no. 2, pp. 60–71, Feb. 2011.
[106] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*. Nelson Education Limited, 2008.
[107] P. Viola and M. J. Jones, “Robust Real-Time Face Detection,” *Int J Comput Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.
[108] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, 2005, vol. 1, pp. 886–893 vol. 1.
[109] J. Schlenzig, E. Hunter, and R. Jain, “Recursive identification of gesture inputs using hidden Markov models,” in *, Proceedings of the Second IEEE Workshop on Applications of Computer Vision, 1994*, 1994, pp. 187–194.
[110] Y. Kuno, M. Sakamoto, K. Sakata, and Y. Shirai, “Vision-based human interface with user-centered frame,” in *Proceedings of the IEEE/RSJ/GI International Conference on Intelligent Robots and Systems ’94. “Advanced Robotic Systems and the Real World”, IROS ’94*, 1994, vol. 3, pp. 2023–2029 vol.3.
[111] A. K. Bourke, J. V. O’Brien, and G. M. Lyons, “Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm,” *Gait Posture*, vol. 26, no. 2, pp. 194–199, Jul. 2007.
[112] S. K. Pal and S. Mitra, *Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1999.

[113] C. J. C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” *Data Min. Knowl. Discov.*, vol. 2, pp. 121–167, 1998.

[114] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257 –286, Feb. 1989.

[115] J. Yamato, J. Ohya, and K. Ishii, “Recognizing human action in time-sequential images using hidden Markov model,” in *, 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1992. Proceedings CVPR ’92*, 1992, pp. 379 –385.

[116] S. Z. Li, *Markov Random Field Modeling in Computer Vision*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1995.

[117] L. Cruz, D. Lucio, and L. Velho, “Kinect and RGBD Images: Challenges and Applications,” in *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, 2012, pp. 36 –49.

[118] B. Yoo, J.-J. Han, C. Choi, K. Yi, S. Suh, D. Park, and C. Kim, “3D user interface combining gaze and hand gestures for large-scale display,” in *CHI ’10 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2010, pp. 3709–3714.

[119] F. Klompmaker, K. Nebe, and A. Fast, “dSensingNI: a framework for advanced tangible interaction using a depth camera,” in *Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction*, New York, NY, USA, 2012, pp. 217–224.

[120] C. Bellmore, R. Ptucha, and A. Savakis, “Interactive display using depth and RGB sensors for face and gesture control,” in *Image Processing Workshop (WNYIPW), 2011 IEEE Western New York*, 2011, pp. 1 –4.

[121] Jacob, Li, and Wachs, “Gestonurse: a multimodal robotic scrub nurse,” in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, New York, NY, USA, 2012, pp. 153–154.

[122] D. Droeschel, J. Stückler, and S. Behnke, “Learning to interpret pointing gestures with a time-of-flight camera,” in *Proceedings of the 6th international conference on Human-robot interaction*, New York, NY, USA, 2011, pp. 481–488.

[123] M. Van den Bergh, D. Carton, R. de Nijs, N. Mitsou, C. Landsiedel, K. Kuehnlenz, D. Wollherr, L. Van Gool, and M. Buss, “Real-time 3D hand gesture interaction with a robot for understanding directions from humans,” in *2011 IEEE RO-MAN*, 2011, pp. 357 –362.

[124] K. R. Konda, A. Königs, H. Schulz, and D. Schulz, “Real time interaction with mobile robots using hand gestures,” in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, New York, NY, USA, 2012, pp. 177–178.

[125] C. Keskin, F. Kirac, Y. E. Kara, and L. Akarun, “Real time hand pose estimation using depth sensors,” in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 1228 –1234.

[126] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, “American sign language recognition with the kinect,” in *Proceedings of the 13th international conference on multimodal interfaces*, New York, NY, USA, 2011, pp. 279–286.

[127] H.-K. Lee and J. H. Kim, “An HMM-based threshold model approach for gesture recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 961–973, Oct. 1999.

[128] S. S. Rautaray and A. Agrawal, “Vision based hand gesture recognition for human computer interaction: a survey,” *Artif. Intell. Rev.*, pp. 1–54, Nov. 2012.

[129] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, “An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition,” *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1035–1074, Apr. 1983.

[130] R. P. Lippmann, “Review of Neural Networks for Speech Recognition,” *Neural Comput*, vol. 1, no. 1, pp. 1–38, Mar. 1989.

[131] K.-F. Lee and H.-W. Hon, “Large-vocabulary speaker-independent continuous speech recognition using HMM,” in *, 1988 International Conference on Acoustics, Speech, and Signal Processing, 1988. ICASSP-88*, 1988, pp. 123–126 vol.1.

[132] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, and D. Ollason, “HTKBook.” [Online]. Available: http://htk.eng.cam.ac.uk/.

[133] M. J.-D. Otis and B. J. Menelas, “Toward an augmented shoe for preventing falls related to physical conditions of the soil,” in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2012, pp. 3281–3285.

[134] S. Brassard, M. J.-D. Otis, A. Poirier, and B.-A. J. Menelas, “Towards an Automatic Version of the Berg Balance Scale Test Through a Serious Game,” in *Proceedings of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare*, New York, NY, USA, 2012, pp. 5:1–5:6.

[135] J. J. LaViola Jr., D. A. Feliz, D. F. Keefe, and R. C. Zeleznik, “Hands-free Multi-scale Navigation in Virtual Environments,” in *Proceedings of the 2001 Symposium on Interactive 3D Graphics*, New York, NY, USA, 2001, pp. 9–15.

[136] E. Fassbender and D. Richards, “Using a Dance Pad to Navigate through the Virtual Heritage Environment of Macquarie Lighthouse, Sydney,” in *Virtual Systems and Multimedia*, T. G. Wyeld, S. Kenderdine, and M. Docherty, Eds. Springer Berlin Heidelberg, 2008, pp. 1–12.

[137] J. A. Paradiso, S. J. Morris, A. Y. Benbasat, and E. Asmussen, “Interactive Therapy with Instrumented Footwear,” in *CHI ’04 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2004, pp. 1341–1343.

[138] F. Daiber, J. Schöning, and A. Krüger, “Whole Body Interaction with Geospatial Data,” in *Smart Graphics*, A. Butz, B. Fisher, M. Christie, A. Krüger, P. Olivier, and R. Therón, Eds. Springer Berlin Heidelberg, 2009, pp. 81–92.

[139] T. Pakkanen and R. Raisamo, “Appropriateness of Foot Interaction for Non-accurate Spatial Tasks,” in *CHI ’04 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2004, pp. 1123–1126.

[140] J. Schöning, F. Daiber, A. Krüger, and M. Rohs, “Using Hands and Feet to Navigate and Manipulate Spatial Data,” in *CHI ’09 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2009, pp. 4663–4668.

[141] F. Daiber, J. Schöning, and A. Krüger, “Towards a Framework for Whole Body Interaction with Geospatial Data,” in *Whole Body Interaction*, D. England, Ed. Springer London, 2011, pp. 197–207.

[142] F. Göbel, K. Klamka, A. Siegel, S. Vogt, S. Stellmach, and R. Dachselt, "Gaze-supported Foot Interaction in Zoomable Information Spaces," in *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2013, pp. 3059–3062.

[143] M. A. Pérez-Quiñones and J. L. Sibert, "A collaborative model of feedback in human-computer interaction," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 1996, pp. 316–323.

[144] G. Jefferson, "Notes on a systematic deployment of the acknowledgement tokens 'Yeah'; and 'Mm Hm';," *Pap. Linguist.*, vol. 17, no. 2, pp. 197–216, 1984.

[145] S. J. Sigman, *The Consequentiality of Communication*. Psychology Press, 1995.

[146] B.-C. Bae, A. Brunete, U. Malik, and E. Dimara, "Towards an Empathizing and Adaptive Storyteller System," *Eighth Artif. Intell. Interact. Digit. Entertain. Conf.*, 2012.

[147] F. He and A. Agah, "Multi-Modal Human Interactions with an Intelligent Interface Utilizing Images, Sounds, and Force Feedback," *J Intell Robot. Syst*, vol. 32, no. 2, pp. 171–190, Oct. 2001.

[148] J. Pasquero and V. Hayward, "Tactile feedback can assist vision during mobile interactions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2011, pp. 3277–3280.

[149] X. Zhang, W. Feng, and H. Zha, "Effects of Different Visual Feedback Forms on Eye Cursor's Stabilities," in *Internationalization, Design and Global Development*, P. L. P. Rau, Ed. Springer Berlin Heidelberg, 2011, pp. 273–282.

[150] H. Richter, A. Hang, and B. Blaha, "The PhantomStation: towards funneling remote tactile feedback on interactive surfaces," in *Proceedings of the 2nd Augmented Human International Conference*, New York, NY, USA, 2011, pp. 5:1–5:2.

[151] T. Stockinger and H. Richter, "Multi-Haptics: Remote Tactile Feedback on Multitouch Surfaces," *Proc. Mensch Comput. 2012*, Sep. 2012.

[152] A. W. Salmoni, R. A. Schmidt, and C. B. Walter, "Knowledge of results and motor learning: A review and critical reappraisal," *Psychol. Bull.*, vol. 95, no. 3, pp. 355–386, 1984.

[153] C. J. Winstein, "Knowledge of results and motor learning--implications for physical therapy," *Phys. Ther.*, vol. 71, no. 2, pp. 140–149, Feb. 1991.

[154] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Mach. Learn.*, vol. 20, no. 3, pp. 197–243, Sep. 1995.

[155] F. V. Jensen, *Bayesian Networks and Decision Graphs*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2001.

[156] D. Heckerman, "A Tutorial on Learning with Bayesian Networks," in *Innovations in Bayesian Networks*, P. D. E. Holmes and P. L. C. Jain, Eds. Springer Berlin Heidelberg, 2008, pp. 33–82.

[157] A. J. Yu and P. Dayan, "Inference, attention, and decision in a Bayesian neural architecture," in *Advances in Neural Information Processing Systems 17*, 2005, pp. 1577–1584.

[158] R. P. N. Rao, “Bayesian inference and attentional modulation in the visual cortex,” *Neuroreport*, vol. 16, no. 16, pp. 1843–1848, Nov. 2005.
[159] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, “What and where: a Bayesian inference theory of attention,” *Vision Res.*, vol. 50, no. 22, pp. 2233–2247, Oct. 2010.
[160] C. Conati and X. Zhao, “Building and evaluating an intelligent pedagogical agent to improve the effectiveness of an educational game,” in *Proceedings of the 9th international conference on Intelligent user interfaces*, New York, NY, USA, 2004, pp. 6–13.
[161] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, *A Bayesian Approach to Filtering Junk E-Mail*. 1998.
[162] E. Horvitz, J. Breese, D. Heckerman, D. Hovel, and K. Rommelse, “The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users,” in *In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 1998, pp. 256–265.
[163] S. Gievska and J. Sibert, “Using task context variables for selecting the best timing for interrupting users,” in *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*, New York, NY, USA, 2005, pp. 171–176.
[164] G. F. Cooper and E. Herskovits, “A Bayesian method for the induction of probabilistic networks from data,” *Mach. Learn.*, vol. 9, no. 4, pp. 309–347, Oct. 1992.
[165] J. Rissanen, “Stochastic Complexity and Modeling,” *Ann. Stat.*, vol. 14, no. 3, pp. 1080–1100, Sep. 1986.
[166] R. R. Bouckaert, “Probabilistic network construction using the minimum description length principle,” in *Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, M. Clarke, R. Kruse, and S. Moral, Eds. Springer Berlin Heidelberg, 1993, pp. 41–48.
[167] D. D. R. Brodbeck and S. E. Tanninen, “Place Learning and Spatial Navigation,” in *Encyclopedia of the Sciences of Learning*, P. D. N. M. Seel, Ed. Springer US, 2012, pp. 2639–2641.
[168] J. Bures, O. Buresová, and L. Nerad, “Can rats solve a simple version of the traveling salesman problem?,” *Behav. Brain Res.*, vol. 52, no. 2, pp. 133–142, Dec. 1992.
[169] T. Tenbrink and J. Wiener, “The verbalization of multiple strategies in a variant of the traveling salesperson problem,” *Cogn. Process.*, vol. 10, no. 2, pp. 143–161, May 2009.
[170] J. N. MacGregor, E. P. Chronicle, and T. C. Ormerod, “Convex hull or crossing avoidance? Solution heuristics in the traveling salesperson problem,” *Mem. Cognit.*, vol. 32, no. 2, pp. 260–270, Mar. 2004.
[171] D. Vickers, M. D. Lee, M. Dry, P. Hughes, and J. A. McMahon, “The aesthetic appeal of minimal structures: Judging the attractiveness of solutions to traveling salesperson problems,” *Percept. Psychophys.*, vol. 68, no. 1, pp. 32–42, Jan. 2006.
[172] A. Blum, S. Chawla, D. R. Karger, T. Lane, A. Meyerson, and M. Minkoff, “Approximation Algorithms for Orienteering and Discounted-Reward TSP,” *SIAM J Comput*, vol. 37, no. 2, pp. 653–670, May 2007.

[173] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *In In CVPR, 2011. 3*.
[174] A. Johnson and R. W. Proctor, *Attention: Theory and Practice*. SAGE, 2004.
[175] H. E. Pashler, *The Psychology of Attention*. MIT Press, 1999.
[176] K. B. Korb and A. E. Nicholson, *Bayesian Artificial Intelligence*. CRC Press, 2003.
[177] P. Venkataraman, *Applied Optimization with MATLAB Programming*. John Wiley & Sons, 2002.
[178] P. Larranaga, M. Poza, Y. Yurramendi, R. H. Murga, and C. M. H. Kuijpers, "Structure learning of Bayesian networks by genetic algorithms: a performance analysis of control parameters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 9, pp. 912–926, Sep. 1996.
[179] N. Friedman, "Learning Belief Networks in the Presence of Missing Values and Hidden Variables," in *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, pp. 125–133.
[180] R. E. Neapolitan, *Learning Bayesian networks*. Upper Saddle River, N.J.: Pearson Prentice Hall, 2004.
[181] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Commun ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.
[182] Z. Zhao and J. Shah, "A Normative DFM Framework Based on Benefit-Cost Analysis," *Proc. DETC 2002 ASME 2002 Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf.*, pp. 227–238, Jan. 2002.
[183] Y.-T. Li and J. P. Wachs, "HEGM: A hierarchical elastic graph matching for hand gesture recognition," *Pattern Recognit.*
[184] M. R. K. Baumann, D. Rösler, and J. F. Krems, "Situation Awareness and Secondary Task Performance While Driving," in *Engineering Psychology and Cognitive Ergonomics*, D. Harris, Ed. Springer Berlin Heidelberg, 2007, pp. 256–263.
[185] L. Foglia and R. A. Wilson, "Embodied cognition," *Wiley Interdiscip. Rev. Cogn. Sci.*, vol. 4, no. 3, pp. 319–325, May 2013.
[186] A. D. Wilson and S. Golonka, "Embodied cognition is not what you think it is," *Cogn. Sci.*, vol. 4, p. 58, 2013.

APPENDIX

APPENDIX QUESTIONNAIRE

Quantitative measures. 1). *Usability*: User ratings of comfort, ease of use and additional human-centered measures for the interface will be collected using a Likert 5 point scale (1 = very hard, 5 = very easy). The subjects will rate several features of image navigation and manipulation control, reflecting the level of suitability to the user. Table 1 below shows an example of the rating scale and questionnaire used.

Table 1. Usability questionnaire.

Rate the following features for user-centered functionality on a scale of 1 to 5:

1. How **comfortable** was the type of command for the given navigational task?

1	2	3	4	5
Very uncomfortable	Uncomfortable	Moderately comfortable	Comfortable	Very comfortable

2. How **precise** was the navigational control for delivering commands?

1	2	3	4	5
Very ambiguous	Ambiguous	Moderately precise	Precise	Very precise

3. How **easy** was the use of command to issue the navigational task?

1	2	3	4	5
Very easy	Easy	Moderately easy	Hard	Very Hard

4. How **frustrating** was the use of command for the navigational task?

1	2	3	4	5
Very pleasant	Pleasant	Moderately frustrated	Frustrated	Very frustrated

5. How **comfortable** was the manner in which feedback presented combined with the use of command?

1	2	3	4	5
Very uncomfortable	Uncomfortable	Moderately comfortable	Comfortable	Very comfortable

6. How **helpful** was the feedback combined with the use of command to the decision making for the navigational task?

1	2	3	4	5
Very unhelpful	Unhelpful	Moderately helpful	Helpful	Very helpful

7. How **clear** was the feedback presented?

1	2	3	4	5
Very confusing	Confusing	Moderately clear	Clear	Very clear

8. How **distracting** was the feedback combined with the use of command to the decision making for the navigational task?

1	2	3	4	5
Very low	Low	Moderate	High	Very high

2). *Background*: User's background information is also collected to reflect the demographic distribution of the group. Table 2 shows an example of questionnaire used.

Table 2. Background questionnaire.

1. What is your age?
Under 18 years old 18 - 24 years old 25 - 34 years old 35 – 44 years old Above 45 years old
2. What is your gender?
Male Female

VITA

VITA

Yu-Ting Li
School of Industrial Engineering, Purdue University

Education
B.S., Communications Engineering, 2007, National Chiao-Tung University, Hsin-Chu, Taiwan
M.S., Operations Management, 2010, National Taiwan University, Taipei, Taiwan
Ph.D., Industrial Engineering, 2014, Purdue University, West Lafayette, Indiana

Research Interests
Human-machine interaction,
Pattern recognition
Machine learning

PUBLICATIONS

PUBLICATIONS

[1] Yu-Ting Li, Juan P. Wachs, "Linking Attention to Physical Action in Complex Decision Making Problems" In Proceedings of the 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), October 2014 (to appear).

[2] Yu-Ting Li, Juan P. Wachs, “A Bayesian Approach to Determine Focus of Attention in Spatial and Time-Sensitive Decision Making Scenarios ” in Proceedings AAAI-14 Workshop on Cognitive Computing for Augmented Human Intelligence, Québec City, Québec, Canada, 2014.

[3] Yu-Ting Li, Juan P. Wachs, "HEGM: A Hierarchical Elastic Graph Matching for Hand Gesture Recognition" in Pattern Recognition (PR), vol. 47, no. 1, pp. 80-88, Jan., 2014.

[4] Yu-Ting Li, Juan P. Wachs, "Recognizing Hand Gestures using the Weighted Elastic Graph Matching (WEGM) Method" in Image and Vision Computing, vol. 31, no. 9, pp. 649-657, Sep. 2013

[5] Yu-Ting Li, Mithun George Jacob, Juan P. Wachs, "A Cyber-Physical Management System for Delivering and Monitoring Surgical Instruments in the OR" in Surgical Innovation, vol. 20, no. 4, pp. 377-384, Aug. 2013.

[6] Mithun George Jacob, Yu-Ting Li, Juan P. Wachs, George Akingba, "Collaboration with a Robotic Scrub Nurse" in Communications of the ACM (CACM), vol. 56, no. 6, pp. 68-75, May 2013.

[7] Mithun George Jacob, Yu-Ting Li, Juan P. Wachs, "Surgical Instrument Handling and Retrieval in the Operating Room with a Multimodal Robotic Assistant" in the Proceedings of the 2013 IEEE International Conference on Robotics and Automation (ICRA), May 6-10 2013, Karlsruhe, Germany

[8] Yu-Ting Li, Juan P. Wachs, "Hierarchical Elastic Graph Matching for Hand Gesture Recognition" in Proceedings of the 16th Iberoamerican Congress conference on Progress in Pattern Recognition (CIARP), Image Analysis, Computer Vision, and Applications, Buenos Aires, Argentina, 2012, pp. 308-315. [9] Mithun George Jacob, Yu-Ting Li, Juan P. Wachs, "Gestonurse: a multimodal robotic scrub nurse" in Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI), Boston, MA, March 2012, pp.153-154.

[10] Juan P. Wachs, Mithun George Jacob, Yu-Ting Li, A. George Akingba, "Does a robotic scrub nurse improve economy of movements?" in Proceeding of SPIE 8316,

Medical Imaging 2012: ImageGuided Procedures, Robotic Interventions, and Modeling, 83160E, 2012

[11] Mithun George Jacob, Yu-Ting Li, Juan P. Wachs, "Gestonurse: a robotic surgical nurse for handling surgical instruments in the operating room" in Journal of Robotic Surgery, vol. 6, no. 1, pp. 53-63, March 2012.

[12] Mithun George Jacob, Yu-Ting Li, Juan P. Wachs, "A Gesture Driven Robotic Scrub Nurse" In Proceedings of the 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2011, pp. 2039-2044.